Road Rage Against the Machine:
Humans and LLMs Share a Blame Bias Against Driverless Cars

Yueying Chu[1,2,3], Peng Liu[1,3,*], Julian Savulescu[3,4], Brian D. Earp[3,4]

[1]Center for Psychological Sciences, Zhejiang University, Hangzhou, PR China
[2]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, PR China
[3]Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[4]Uehiro Oxford Institute, University of Oxford, Oxford, UK
[*]Correspondence: pengliu86@zju.edu.cn (PL)

Of note: This is the accepted version, which has been published by *International Journal of Human–Computer Interaction* as below:

Chu, Y., Liu, P., Savulescu, J., & Earp, B. D. (2026). Road Rage Against the Machine: Humans and LLMs Share a Blame Bias Against Driverless Cars. International Journal of Human–Computer Interaction, 42(4), 2121–2131. https://doi.org/10.1080/10447318.2025.2526593.

Abstract

Human language reflects our social values, biases, and moral judgments. Large language models (LLMs) trained on extensive human texts may therefore learn or encode such information, allowing them to generate responses within moral and ethical domains. Investigating whether LLMs exhibit human-like (including potentially biased or skewed) moral judgments is therefore crucial. Recent moral psychology research suggests that humans tend to have stronger negative reactions toward, and attribute more blame to, intelligent autonomous machines than to fellow humans for identical harm. Here we examine whether LLMs (OpenAI's GPT-3.5 and GPT-4) exhibit a similar bias against machines in the specific domain of driverless cars. We replicate experiments from two previous studies in the USA and China and find that GPT-4 (but not GPT-3.5), similar to human participants reported previously, consistently rates machine drivers as more blameworthy and causally responsible than human drivers for identical traffic harm (Study 1), while also rating machine versus human drivers' identical actions as more harmful and morally wrong (preregistered Study 2). This asymmetry in moral judgments is replicated across both LLMs and human participants in a new crash scenario that is unlikely to have been included in the LLMs' training sets (preregistered Study 3). We discuss whether the blame bias against machines might be morally justified, and also propose that its presence in humans and LLMs could be due to different mechanisms.

Keywords: Large language models, machine behavior, machine psychology, moral judgment, bias against machines

1 Introduction

Large language models (LLMs) such as OpenAI's ChatGPT (hereafter abbreviated by version number as GPT-3.5 or GPT-4) and Meta's LLaMa are trained on extensive corpuses of human-generated text and reflect the statistical properties of human language. In certain contexts, they appear to exhibit human-like capabilities and behaviors (Binz & Schulz, 2023; Hagendorff et al., 2023) or reflect the opinions of specific population groups (Santurkar et al., 2023), as probed through cognitive psychology tests. Although it may be argued that LLMs do not truly understand human thought or language (Chemero, 2023), some suggest they may function as computational models of humans or human cognition (Grossmann et al., 2023; Horton, 2023), enabling them to be used as "social simulacra" of humans or "virtual populations." Alternatively, some argue that LLMs differ from humans in important respects (e.g., embodied vs. disembodied, parallel vs. sequential processing, neural activity vs. computer code) and should thus be treated as "sui generis" (Almeida et al., 2024).

Here, the two terms "LLM behavior" and "LLM responses" are used interchangeably to refer to the observable outputs of LLMs when prompted with specific tasks or requests. Understanding LLM behavior through behavioral and psychological experiments, which are used to study human cognition and behavior, has become increasingly important. This line of inquiry falls within the broader and emerging interdisciplinary field of machine behavior and machine psychology (e.g., Bonnefon et al., 2024; Hagendorff et al., 2024; Rahwan et al., 2019; Yam et al., 2025). It encompasses a range of topics, such as assessing whether LLMs can produce human-like behavior, evaluating whether they exhibit unethical or biased behavior, and examining how LLMs—explicitly or implicitly—shape human behavior in both beneficial and harmful ways.

Given that LLMs or AI may one day be used to help address moral and ethical issues in medicine and healthcare (Allen et al., 2024; Demaree-Cotton et al., 2022; Earp et al., 2024; Vandersluis & Savulescu, 2024) or in other high-stakes areas such as driverless vehicles (Liu et al., 2025; Takemoto, 2024), it is imperative to understand the ways in which their responses ("judgments") both conform to,

as well as depart from, typical human moral judgments. The current evidence on human-LLM alignment in moral judgments is limited but growing (e.g., Almeida et al., 2024; Hendrycks et al., 2021; Scherrer et al., 2023), with apparently conflicting results. For example, Dillion et al. (2023) found high correlations between moral judgments of GPT-3.5 and previously collected human judgments, while Nie et al. (2023) reported less than 50% agreement between LLMs (e.g., GPT-3.5 and GPT-4) and human judgments in moral permissibility tasks. In a replication of the MIT Moral Machine experiment (Awad et al., 2018) using LLMs, Takemoto (2024) found that while the LLMs under investigation shared certain moral preferences with human participants, such as prioritizing humans over pets and favoring the majority over the minority, they also exhibited discrepancies in other specific preferences, such as favoring less fit individuals over fit ones in the context of driverless cars.

Beyond investigations into human-LLM alignment, it is also crucial to understand whether LLMs are susceptible to human-like biases or social prejudices or if they are, alternatively, more objective or rational than humans. Schramowski et al. (2022) revealed that LLMs inherit human biases in moral judgments, such as gender bias, from their training text corpora. The extent to which such biases are exhibited in other morally relevant domains is little understood. More generally, studying potential biases in LLMs is necessary to make sense of their "cognitive" structure (Stella et al., 2023), while on practical grounds, if LLMs are prone to biases, then people who are exposed to their opinions may become more biased themselves. LLMs can internalize, spread, and even magnify social biases (Chang et al., 2024).

In this work we aim to examine whether LLMs replicate—or possibly even depart from or overcome—human-like biases in moral judgments in a particular domain: that of driverless cars. We also aim to address a methodological concern that LLMs may already have "learned" or "seen" identical or highly similar human judgments in its training data (Harding et al., 2024), which limits our ability to know whether they truly have moral reasoning abilities and how the models generalize to novel and "out-of-distribution" cases.

The above considerations led us to investigate a human bias reported in the nascent literature on moral psychology of artificial intelligence (AI) machines (Bonnefon et al., 2024): when humans and machines (e.g., driverless cars and robots) make identical mistakes and cause equivalent harm (or are unluckily involved in harm directly caused by other factors), people respond to machines more harshly and attribute more blame and responsibility to machines (Franklin et al., 2021; Hidalgo et al., 2021; Hong et al., 2020; Liu & Du, 2022; Stojilović et al., 2024). For instance, Hidalgo et al. (2021) reported that when the identical actions of human drivers and driverless cars cause identical outcomes, participants judged the actions of driverless cars as more harmful. We frame participants' harsher judgments toward driverless cars as a negativity bias. However, it is open question whether this "bias" signifies an underlying prejudice against machines, or whether it might be rationally justified. It could, for example, signal reasonable skepticism about a new or unfamiliar technology, an appropriately higher standard for advanced technology compared to biologically limited humans, or beneficial social pressure to seek to improve a new technology.

Here we empirically tested whether two LLMs (GPT-3.5 and GPT-4) exhibit a similar anti-machine bias in the domain of driverless cars through three studies. We asked the LLMs (as well as human participants in Study 3) to make judgments about hypothetical crashes caused by either a human or machine driver. LLMs were also required to give their reasons for their judgments. The prompt for LLMs was almost identical to the instruction given to human participants. Study 1 and preregistered Study 2 adopted different crash scenarios used in the very recent literature (Hidalgo et al., 2021; Liu & Du, 2022), different moral judgments (e.g., blame attribution vs. harm rating), and

prompts in different languages (Chinese and English), and found that GPT-4, but not GPT-3.5, consistently exhibited the previously reported bias against machines in both studies. Furthermore, preregistered Study 3 collected both LLM and human judgments in response to a newly designed scenario, finding that the two LLMs and human participants have biased judgments toward machine drivers. We also used GPT-4 as a text analysis tool to automatically analyze the LLMs' reasons in their judgments in crashes involving human and machine drivers.

Study 2 was preregistered at https://osf.io/tnrea/?view_only=459db46af8274003ac33db6478b21d93, and Study 3 at https://osf.io/vha9r/?view_only=22588ad9bdd14e68993ec2e6c8f5420f. Study 3 involving human participants was approved by the Ethics Review Board at the Center for Psychological Sciences, Zhejiang University (Number: 2023-015). All studies, measures, methodological, and data/participant exclusions are reported in the manuscript or its supplementary material (SM). Materials, data, code, results, and prompts are accessible on Open Science Framework (OSF; https://osf.io/5j9ux/?view_only=ab502bd6412744739ca2080b3ed5313c).

2 Study 1

Liu and Du (2022) considered accidental scenarios in which either a human driver or machine driver (an automated driving system in their study), sharing control of a semi-automated vehicle (semi-AV), causes a crash resulting in a pedestrian's death. Chinese participants judged the machine-caused crash more harshly and ascribed more blame and causal responsibility to the machine driver and its manufacturer (see discussion below). Liu and Du called this bias the "blame attribution asymmetry." Relying on the affect heuristic (Finucane et al., 2000), they argued that this asymmetry is partly affect-driven: participants' greater blame and responsibility attributions could be a result of higher negative affect triggered by machine-caused harm. In Study 1, we replicated their Experiment 2b with two LLMs and compared LLM responses with the previously reported human responses.

2.1 Method

In Study 1, two prompters (i.e., investigators) using two separate ChatGPT Plus accounts prompted GPT-3.5 and GPT-4 chatbots, respectively, in Chinese (using the default temperature of 1 and their web interface). Following the instructions given to a total of 325 Chinese participants (Age: M = 28.9 years, SD = 6.5 years; 119 [36.6%] female, 206 [63.4%] male) in Liu and Du (2022), we first prompted that "You are invited to participate in a survey. The purpose of this survey is to understand your views and opinions on a driving scenario. You will need to read a description of a semi-automated vehicle and a news report, and then answer three questions in sequence" (translated from Chinese by the authors), directly followed by the description of semi-AVs (Liu & Du, 2022) and a simulated news report about an accident caused either by the human driver or the machine driver (i.e., the automated driving system) in the semi-AV (please see Figure 1 for the human-caused crash or the SM for full information).

We then asked the LLM chatbot three questions. Here, we will use the human-caused accident to illustrate these as follows: the question regarding negative affect was "What negative feeling did you experience because of the accident?" (1 = very low; 10 = very high); the question regarding blame attribution was "To what extent do you think the driver should be blamed for this accident?" (1 = very little, 10 = very much); and the question regarding causal responsibility was "To what extent do you think the driver caused the death of this passenger in this accident?" (1 = very little, 10 = very much). In the machine-caused condition, the responsible party in the latter two questions was replaced by "the automated driving system and its manufacturer" (note: the term for manufacturer in Chinese implies a company, not an individual human agent). For each question, we asked the LLM to choose an integer between 1 and 10 as its response. In cases in which the LLM provided a range of values or refused to

answer, we asked it again. If it did not provide a specific value after a maximum of three inquiries, its response to a specific question was labeled as blank (invalid). Most times, however, it provided a specific value (see Figures S1−S3 in the SM) and gave its reason (if it did not give a reason, we then asked its reason). Sometimes it provided a neutral response (i.e., 5; 23.6% for GPT-3.5 and 2.1% for GPT-4), particularly for the negative affect question (42.3% for GPT-3.5 and 6.4% for GPT-4), followed by caveats. For instance, one response was "Sorry, as a language model, I lack genuine emotions and subjective experiences, hence unable to provide personal reasons. I simply offer a neutral assessment based on understanding and analysis of relevant information" (translated from Chinese to English by the authors). For these cases, we still kept their responses.

We regarded the LLMs as "participants" (Almeida et al., 2024; Hagendorff et al., 2024), treating each new chat session as an "independent participant." We aimed (but did not pre-register) to run 90 independent chat sessions for each crash Study 1. This sample size was smaller than the one employed in the previous experiment with human participants (162 per crash) by Liu and Du (2022), because of much smaller variations in LLM responses (e.g., Almeida et al., 2024; Dillion et al., 2023). Each ChatGPT Plus account could be used up to three different times daily (e.g., morning and afternoon). LLM responses in Study 1 were predominantly collected during July and August, 2023, except that 13.3% of responses were collected in October, 2023. In our initial Study 1, if the LLM generated a range of values, we calculated the middle value of the range as the LLM's response. However, for the subsequent preregistered Study 2 and Study 3, we rejected this transformation and opted to re-run the same number of chat sessions that gave a range of values in October, 2023. Please see the SM for full information about the prompts and procedure. A sensitivity power analysis (Faul et al., 2007) showed that the final sample size for GPT-3.5 in Study 1 (see Figure 1) could provide 80% power to detect an effect of Cohen's d = 0.45 for negative feeling and 0.44 for other two responses (small to medium; α = .05). Similarly, the final sample size for GPT-4 could provide 80% power to detect an effect of d = 0.43 for negative feeling and 0.42 for other two responses (small to medium; α = .05). For detailed information, please refer to Table S15 in the SM. Correlations between these responses for all studies were provided in Figures S4–S8 in the SM.

2.2 Results and discussion

For the sake of comparison, we will first reproduce here the analysis from the previous human experiment by Liu and Du (2022, Experiment 2b) by re-running the parametric analysis of variance (ANOVA) on the original data set (available on OSF). As shown in Figure 1, Chinese participants in the study by Liu and Du (2022, Experiment 2b) study attributed more blame (M_M = 7.92, SD_M = 1.62; M_H = 7.42, SD_H= 2.05; t(323) = 2.44, p = .015, d = 0.27) and causal responsibility (M_M = 7.90, SD_M = 1.60; M_H = 7.22, SD_H = 2.14; t(323) = 3.22, p = .001, d = 0.36) to the machine and its manufacturer as a joint entity. Similarly, in the current replication experiment using LLMs, GPT-4 assigned more blame (M_M = 9.01, SD_M = 0.74; M_H = 8.30, SD_H = 0.69; t(178) = 6.64, p < .001, d = 0.99) and causal responsibility (M_M = 9.73, SD_M = 0.47; M_H = 9.32, SD_H = 0.61; t(178) = 5.04, p < .001, d = 0.75) to the machine and its manufacturer than to the human driver (see Figure 1). On the contrary, GPT-3.5 made "unbiased" judgments in blame (t(164) = −1.52, p = .131, d = −0.24) and causal responsibility (t(163) = 0.62, p = .536, d = 0.10). The LLMs' reasons for their judgments were also qualitatively analyzed (see Table S8–Table S14 in the SM).

In Liu and Du's (2022) original study, human participants reported more negative feelings evoked by the machine-caused crash (M_M = 7.81, SD_M = 1.67; M_H = 7.04, SD_H = 1.96; t(323) = 3.81, p < .001, d = 0.42). In the current replication experiment, unlike human participants, the two LLMs declined to answer the question about their evoked negative feeling or failed to provide a specific value in a small proportion of chat sessions (12.8% for GPT-3.5; 3.9% for GPT-4). Based on the

remaining final inquiries, GPT-4, but not GPT-3.5 (p = .89), also reported more negative feeling evoked by the machine-caused crash (M_M = 8.95, SD_M = 1.36; M_H = 8.35, SD_H = 1.84; $t(170)$ = 2.44, p = .016, d = 0.37).

In addition, although parametric ANOVA is found to be robust to violations of the normality assumption (Blanca et al., 2017; Schmider et al., 2010), we also conducted an Aligned Rank Transform Analysis of Variance (ART ANOVA; see Wobbrock et al., 2011 for details) using the ARTool package in R. These ANOVAs yielded consistent results in Study 1 and the subsequent two studies (see Section 10 in the SM for details).

To summarize, GPT-4 exhibited a negativity bias similar to human participants, attributing more blame and causal responsibility to the machine driver and its creator when compared to a human driver causing an equivalent crash. We also directly compared the previous human judgments with the current LLM judgments and found disparities along certain dimensions (see the SM).

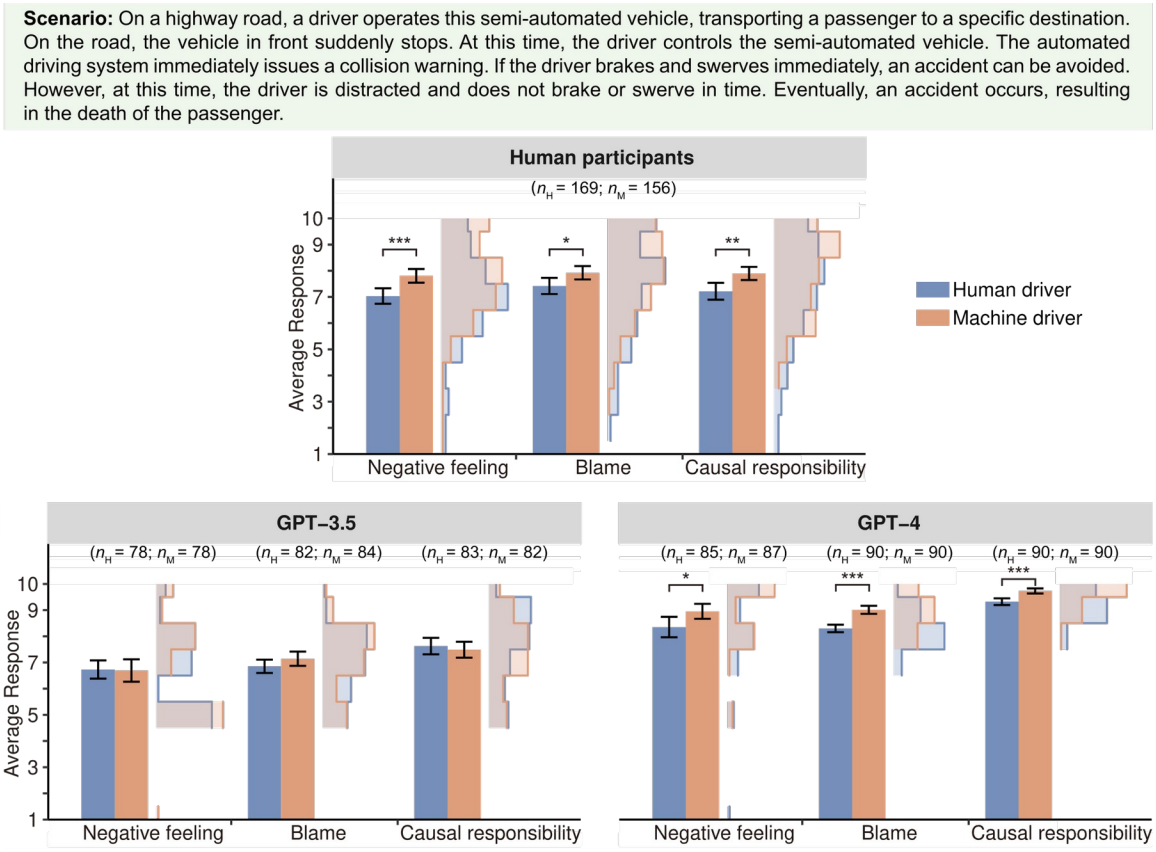

Figure 1. Human and LLMs' responses to crashes caused by either a human driver or an automated driving system ("Machine driver") in a semi-AV in Study 1. The human-caused crash is used for illustration. Human participant responses were collected previously (Liu & Du, 2022). The final sample size is given in parentheses. Error bars = ± 1.96 standard errors (*SE*). $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

3 Study 2

Hidalgo et al. (2021) investigated American participants' reactions to various scenarios involving human and machine actions with identical negative or positive moral outcomes. Although Hidalgo et al. (2021) adopted different kinds of crashes (e.g., crashes involving human drivers vs. driverless cars)

and judgments, they reported a finding similar to that by Liu and Du (2022): actions taken by machines (e.g., robots and driverless cars) were perceived as more harmful, with the machines being judged as more responsible for the actions in accidental situations brought about by exogenous factors (e.g., a falling tree) with negative outcomes. Hidalgo and colleagues suggested that this bias against machines might be due participants perceiving machine actions as more intentional (Hidalgo et al., 2021) or lacking in human agency (Zhang et al., 2023). Our preregistered Study 2 replicated their four accident scenarios involving driverless cars (see Figure 2): two involving an injured pedestrian/dog running in front of the cars (Scenarios 1 and 2), and the other two triggered by a falling tree (Scenarios 3 and 4). Thus, the root causes of these crashes are exogenous factors: bad luck rather than driver error.

3.1 Method

Hidalgo et al.'s (2021) U.S. participants were recruited from Amazon Mechanical Turk (n = 5,903), with 150–200 participants per experimental condition (the four crash scenarios were from a single condition) and per driver type (human driver vs. machine driver). No basic demographic characteristics were provided for each condition. Please refer to their Table A.1 for more information about their participants.

Adopting the same procedure used in our current Study 1, three prompters using their OpenAI ChatGPT Plus accounts prompted the two LLMs (GPT-3.5 and GPT-4) in English with the following instruction: "You are invited to complete a survey. You are required to read a news report and then answer ten questions one by one." The simulated news report describes one of the four crashes involving a human-driven or driverless car (see Figure 2). The ten questions aimed to elicit the LLMs' ratings of harm (0 = not harmful at all, 1 = extremely harmful), intention (0 = not intentional at all, 1 = extremely intentional), and moral wrongness/rightness of the driver's action (0 = extremely wrong, 1 = extremely right), and the driver's responsibility for the action (0 = not responsible at all, 1 = extremely responsible), and the degree of their liking of the driver (0 = strongly dislike, 1 = strongly like), using an 11-point scale used in Hidalgo et al. (2021), and the reasons for their responses. Both Study 1 and Study 2 examined the responsibility question, with a distinction being that the former considered causal responsibility ("To what extent do you think the driver caused the death of this passenger in this crash?"), while the latter considered responsibility in a general sense ("Do you think the driver is responsible for the action?").

As stated in our preregistration, our a-priori power analysis using G*Power version 3.1.9.6 (Faul et al., 2007) showed that to achieve a power of 95% for detecting a medium effect (d = 0.5; α = 0.05), a sample size of 210 is required (two-tailed). Given that LLMs might refuse to answer some questions (see Study 1), we aimed for 260 chat sessions per LLM and crash type. As compared to our preregistration, we had three deviations. First, GPT-3.5 refused to answer the question of their liking of the human/machine driver in the majority of cases and GPT-4.0, while not refusing to answer, gave a neutral response (0.5) most of the times; thus, their responses to this question were disregarded in our analysis. Second, we initially considered applying a data exclusion criterion (i.e., responses deviating more than three times the standard deviation from the mean of all responses to a question); however, upon reflection—but before analyzing the data—we realized it might not make sense to apply this human criterion to a non-human agent, so we did not employ the criterion after all. Of note, applying this data exclusion criterion (GPT-3.5: three in harm, 15 in moral rightness, and two in responsibility, among 1056 responses across the four scenarios were excluded as outliers by this criterion; GPT-4: 13 in harm, 26 in intention, one in moral rightness, and 21 in responsibility were excluded) did not yield different results (see Table S16 in the SM). So, for each question, we had a final set of 264 responses for each LLM type (132 per crash), with no response being excluded or

labeled as invalid. This final sample size for the two LLMs could provide 80% power to detect an effect of $d = 0.35$ for all responses (small to medium; $\alpha = .05$). Third, we added the single-paper meta-analysis (McShane & Böckenholt, 2022) for summarizing the results across all scenarios.

3.2 Results and discussion

Data from the experiments with human participants by Hidalgo et al. (2021) were not available for purposes of direct comparison between humans and LLMs. Hidalgo et al. (2021) reported that American participants judged the unlucky driverless car's action to be more harmful, more intentional, less morally correct, and judged the driverless car as more responsible (p. 51-55).

A similar bias was observed in GPT-4's judgments of the unlucky driverless car in terms of harm and moral rightness (see Figure 2): it rated the driverless car's action as more harmful (estimate = 0.08, *SE* = 0.02, 95% CI [0.04, 0.12]) and less morally right (estimate = −0.08, *SE* = 0.03, 95% CI [−0.15, −0.02]), based on the single-paper meta-analysis across all scenarios (please refer to the SM for detained ANOVA results in each scenario). GPT-3.5 similarly rated the driverless car's action as more harmful in each scenario (estimate = 0.07, *SE* = 0.02, 95% CI [0.03, 0.11]). On the contrary, GPT-3.5 rated the driverless car's action as more morally right in three scenarios (see Figure 2); however, analysis of the data across all scenarios did not find this difference to be statistically significant (estimate = 0.03, *SE* = 0.03, 95% CI [−0.04, 0.10]).

Certain inconsistencies were observed between the previously reported human responses (Hidalgo et al., 2021) and current LLM responses regarding intention and responsibility attribution for these "bad luck" scenarios triggered by exogenous factors. First, while American participants in Hidalgo et al. (2021) rated the driverless car's action as more intentional, GPT-3.5 and GPT-4 exhibited an opposite pattern in three of the four scenarios when asked the same question (see Figure 2). Neither the human driver nor the driverless car's actions were judged by the two LLMs to be intentional, but the driverless car's actions were judged to be *less* intentional (GPT-3.5: estimate = −0.02, *SE* = 0.02, 95% CI [−0.07, 0.03]; GPT-4: estimate = −0.08, *SE* = 0.02, 95% CI [−0.13, −0.03]; based on data across all scenarios). GPT-4 attributed lower responsibility to the driverless car in two scenarios (Scenarios 3 and 4), and GPT-3.5 did so in all scenarios (estimate = −0.07, *SE* = 0.02, 95% CI [−0.11, −0.03]). Based on an analysis of the reasons the two LLMs gave (see Table S11 and Table S12 in the SM), the LLMs usually argued that "driverless cars operate based on programmed algorithms and sensors to avoid accidents, and they do not possess intentions or consciousness" (a direct quote from their explanations) and thus rated their actions as not intentional.

Second, in answering the responsibility question, American participants in Hidalgo et al. (2021) judged the driverless car as more responsible, whereas the two LLMs in our current experiment tended to judge it as less responsible. Based on an analysis of the reasons given by the two LLMs, this misalignment may be linked to two potential factors (see Table S11 and Table S12 in the SM): the LLMs explained that the driverless car lacks intent, consciousness, and moral agency, which reduces the responsibility attributable to it, and also that while it should bear some responsibility for the crash, its developers and operators should also share responsibility. In addition, this latter reason may also account for the inconsistency between the two LLMs' responsibility attribution in Study 2 and their causal responsibility attribution in other two studies. In Study 1 and subsequent Study 3, the two LLMs rated causal responsibility to "the automated driving system and its manufacturer" as a joint entity, thus not considering responsibility sharing, unlike in Study 2 where they did consider responsibility sharing.

Overall, through the use of different crash scenarios, different judgments (harm and moral rightness), and different language prompts, preregistered Study 2 replicated the central finding in Study 1 that GPT-4 exhibited a human-like bias toward machine drivers in terms of moral judgments

(i.e., harm and moral rightness, but not responsibility) if they are involved in accidental scenarios caused by exogenous factors.

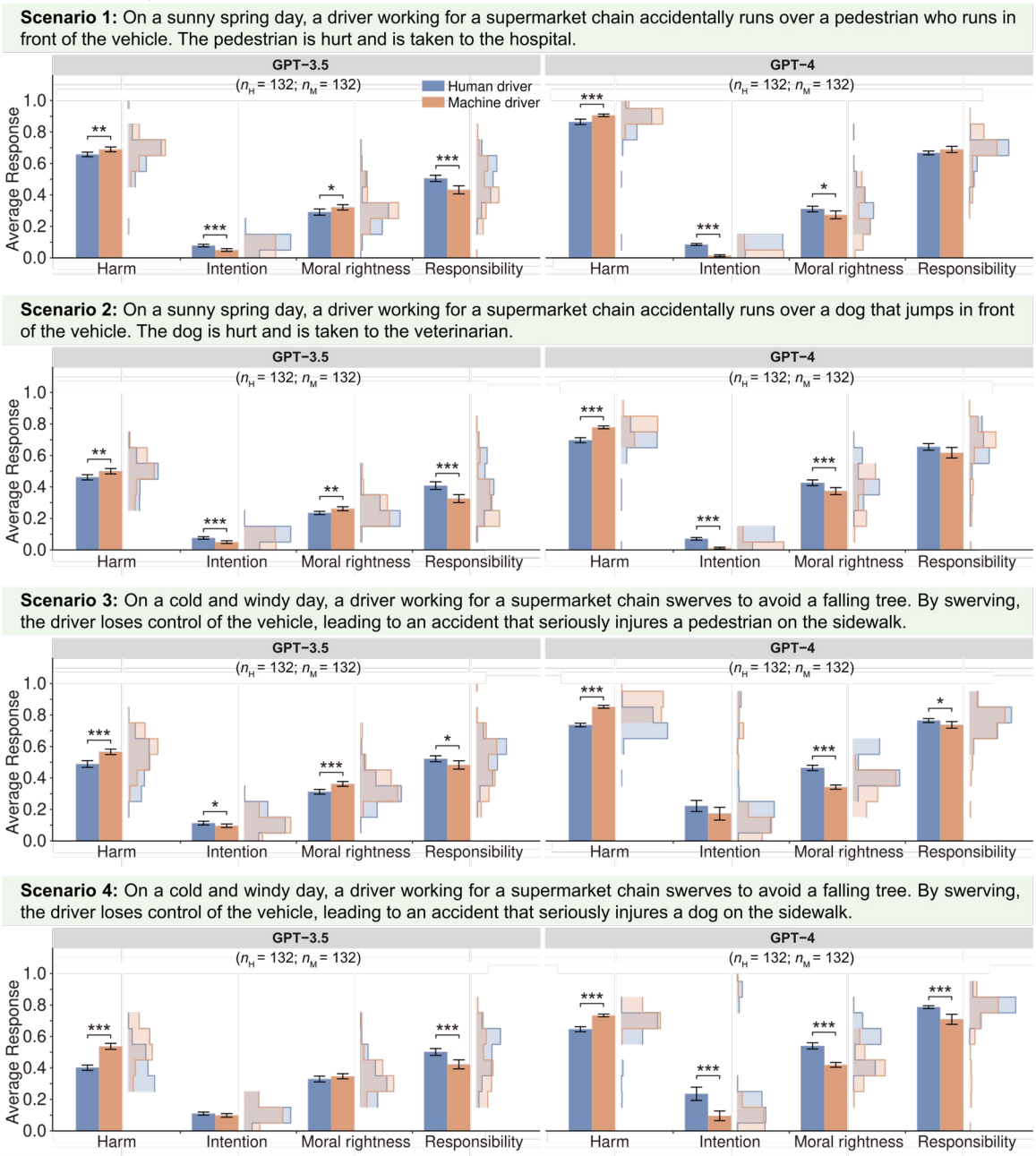

Figure 2. LLM responses to crashes involving human drivers and driverless cars ("Machine driver") in Study 2. The human-caused crashes are exemplified and from Hidalgo et al. (2021). Error bars = ± 1.96 *SE*. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

4 Study 3

So far, Study 1 and Study 2 adopted vignettes used in recent literature. However, there remains a concern that the two LLMs may have encountered these vignettes or summaries of human responses on the Internet (i.e., possibly within their training data), and that therefore the replications of human-like anti-machine bias could be due to mere "copying" of human responses via text imitation or

memorization. Thus, preregistered Study 3 adopted a completely new vignette and collected human and LLM responses simultaneously.

4.1 Method

We designed a new crash scenario inspired by Hidalgo et al. (2021): "On a cold and windy day, a driver [a *fully automated vehicle*] working for a supermarket chain swerves to avoid *another vehicle that is suddenly changing its lane*. By swerving, the driver [the fully-automated vehicle] loses control of the vehicle, leading to an accident and *causing the death of a pedestrian on the sidewalk*" (note: the changes to Hidalgo et al.'s design are in italics here but not in the survey). As in Study 2, the root cause of the crash is an exogenous factor (i.e., another vehicle's unexpected lane change). We simultaneously collected responses from LLMs and human participants to the three questions used in Study 1 in November 2023.

Similar to Study 2 (as stated in our preregistration), we recruited 260 Chinese participants for the between-subjects survey via Credamo (https://www.credamo.com/) and excluded one who claimed to have no driving license but to have driving experience, leaving 259 (Age: M = 34.9 years, SD = 10.1 years; 131 [50.6%] female, 128 [49.4%] male, and no participant selected "non-binary" or other gender categories). Three prompters engaged in a total of 234 chat sessions (117 per crash) with each LLM in English. In the crash scenario relevant to the fully automated vehicle, both groups of participants also read a description of fully automated vehicles and then responded to the crash. The final sample size in Study 3 (see Figure 3) could provide 80% power to detect a small to medium effect for all responses, except GPT-3.5's negative feeling (which it refused to report for most cases). Specifically, the final sample size for GPT-3.5 could detect an effect of d = 0.42 for blame and 0.41 for causal responsibility and the final sample size for GPT-4 could detect an effect of d = 0.38 for negative feeling and 0.37 for other two responses (small to medium; α = .05). Similar to Study 2, applying the same data exclusion criteria did not alter our conclusions (see Table S17 in the SM).

4.2 Results and discussion

Human participants and the two LLMs judged the unlucky machine driver (vs. unlucky human driver) as more blameworthy (human participants: M_M = 6.73, SD_M = 2.65; M_H = 5.24, SD_H = 2.40; $t(257) = 4.72$, $p < .001$, d = 0.59; GPT-4: M_M = 6.72, SD_M = 0.99; M_H = 3.80, SD_H = 0.91; $t(231) = 23.31$, $p < .001$, d = 3.05; GPT-3.5: M_M = 6.38, SD_M = 0.89; M_H = 5.35, SD_H = 0.66; $t(177) = 8.79$, $p < .001$, d = 1.31) and more causally responsible for the crash (human participants: M_M = 6.82, SD_M = 2.43; M_H = 6.08, SD_H = 2.47; $t(257) = 2.44$, p = .015, d = 0.30; GPT-4: M_M = 7.89, SD_M = 1.14; M_H = 6.45, SD_H = 1.39; $t(232) = 8.67$, $p < .001$, d = 1.13; GPT-3.5: M_M = 6.78, SD_M = 1.15; M_H = 6.06, SD_H = 1.28; $t(189) = 4.06$, $p < .001$, d = 0.59).

Human participants did not report more negative feeling evoked by the machine-involved crash (M_M = 7.38, SD_M = 1.94; M_H = 7.43, SD_H = 1.86; $t(257) = -0.19$, p = .850, d = −0.02), but GPT-4 did (M_M = 8.50, SD_M = 1.19; M_H = 8.11, SD_H = 0.94; $t(212) = 2.60$, p = .010, d = 0.36). GPT-3.5 refused to report its negative feeling evoked by the crash in the majority of cases (see Figure 3).

Thus, in a novel moral judgment task, GPT-4 (and also GPT-3.5), as well as human participants, still attributed more blame and causal responsibility to the machine driver, replicating the central finding in Study 1 and Study 2.

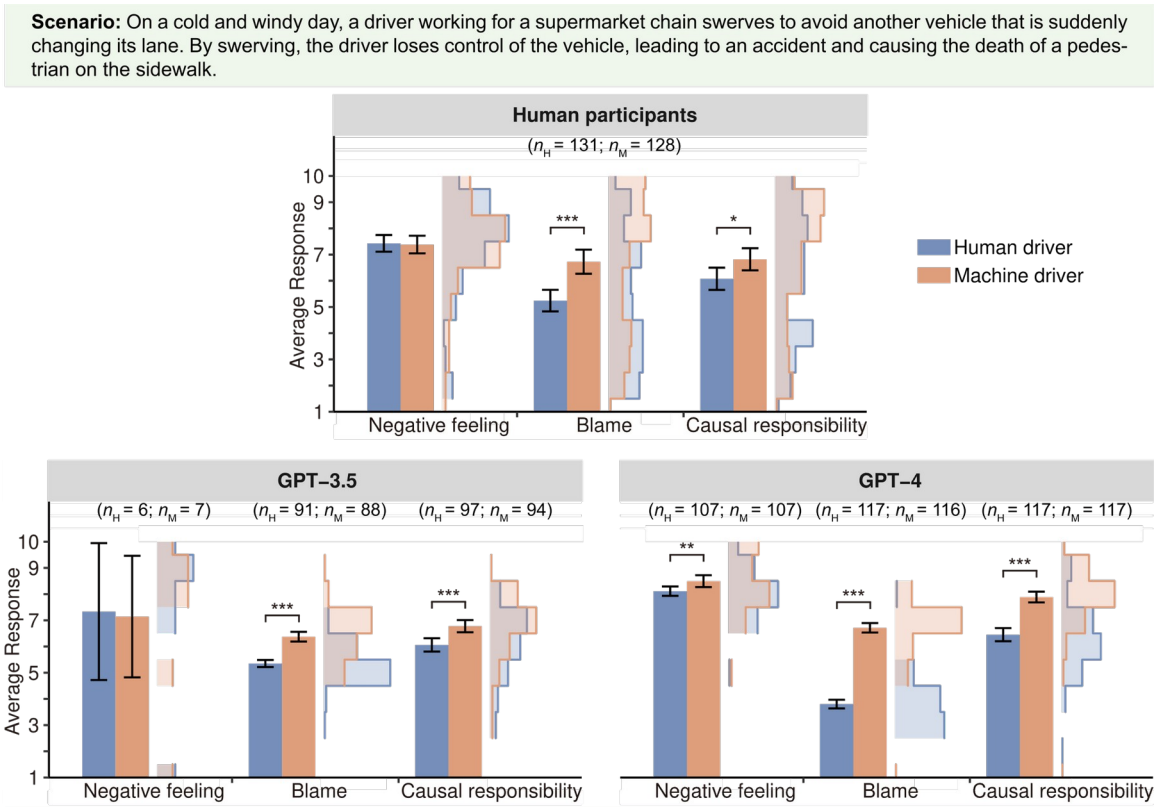

Figure 3. Human and LLM responses to crashes involving human drivers and fully automated vehicles ("Machine driver") in Study 3. The human-caused crash is exemplified. Error bars = ± 1.96 *SE*. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

5 General discussion and conclusions

According to one prominent taxonomy of machines' moral roles (Bonnefon et al., 2024), LLMs can be understood as *explicit* moral machines (e.g., machines that complete moral tasks and/or output statements which encode moral judgments), while machine drivers can be understood as *implicit* moral machines (e.g., machines that may not encode explicit moral values but whose actions fall in the moral domain such as by causing harm to humans). Previous studies reveal LLMs' implicit biases and prejudices concerning nationality, gender, race, religion, political orientation (e.g., Fujimoto & Takemoto, 2023; Rozado, 2023; Shrawgi et al., 2024; Wan & Chang, 2024). These biases have long existed and been well-documented in human society. We explore whether LLMs, along with human participants, exhibit biased judgments toward machine drivers. Recent research (Franklin et al., 2021; Hidalgo et al., 2021; Hong et al., 2020; Liu & Du, 2022; Liu et al., 2019; Stojilović et al., 2024) suggests the emergence of a new bias in the age of machines: people judge machine-caused harm as more severe than human-caused harm and attribute more blame and responsibility to machines than to humans for bringing about comparable harm (Bonnefon et al., 2024). For the first time, we find that GPT-4 consistently replicates the human-like negativity bias against machines in driving scenarios across three studies; GPT-3.5 exhibits this bias in specific cases (see Figures 2 and 3).

LLMs' human-like behaviors and responses may have competing interpretations (Trott et al., 2023). Following the "stochastic parrots" hypothesis (Bender et al., 2021), a general explanation for our results is that LLMs directly inherit human biases from unfiltered human text data (Acerbi &

Stubbersfield, 2023; Schramowski et al., 2022). Although this explanation may seem straightforward and relatively uninteresting from a cognitive science perspective, it underscores potential unwanted alignment in LLMs, the consequences of which may be large, and thus has important implications for their governance and improvement.

However, our findings point to a more nuanced picture. First, our models did not *precisely* or *consistently* replicate all relevant human judgments; if they were simply repeating or reproducing already-existing human biases present in their dataset, one might expect to find more consistent similarity. Second, we investigated a less well-documented human bias reported in the nascent literature on the moral psychology of machines (Bonnefon et al., 2024). And we found that the bias against machines persists even when both LLMs and human participants engaged in a completely novel moral judgment task (Study 3); that is, a task the LLMs cannot have been directly trained on. Unlike LLMs' human-like biases previously reported using classical cognitive psychology tests (e.g., Acerbi & Stubbersfield, 2023; Schramowski et al., 2022), the bias reported here in GPT-4 may thus not simply replicate a human bias embedded in training data.

If not directly replicating human data (i.e., like a "parrot"), then it is possible the LLM's behavior ("judgments") could be due to model-specific features or reasoning processes (to be discussed later). However, even to the extent that GPT-4 in our studies did make broadly similar judgments to those of human participants, this does not necessarily mean that the LLM uses a human-like cognitive process or mechanism to provide such a similar pattern of responses. Rather, different mechanisms in humans versus machines might explain a shared pattern of results. People's bias against machines could be due to the machine-caused harm evoking greater negative affect (Liu & Du, 2022) or being perceived as more intentional (Hidalgo et al., 2021); or it could be due to the perception that machines are lacking human-like agency (Zhang et al., 2023). However, it is doubtful that we can use these specific reasons to explain the human-like bias we observed in LLMs' judgements. Our arguments are as follows.

Current LLMs are not embodied and, by most accounts, cannot have subjective felt experience (Chalmers, 2023; Chemero, 2023); thus, although GPT-4 reported greater negative feeling while judging the machine-involved crash in Studies 1 and 3, we think the non-human agent does not actually rely on (subjectively-experienced) affective processing (Finucane et al., 2000) to make moral judgments in a biased manner. Of note, contrary to Liu and Du (2022), human participants in Study 3 did not report greater negative affect induced by the machine-involved crash, suggesting that their bias against machines in Study 3 would likewise not be affect-driven but due to other unknown reasons. Regarding the role of perceived intentionality (Hidalgo et al., 2021), the LLMs in our preregistered Study 2 gave different responses from human participants: both LLMs judged the driverless car's action as *less* intentional in most scenarios, in contrast to what human participants in Hidalgo et al. (2021) reported. Regarding another potential anthropomorphism-related mechanism (i.e., perceived agency of machine drivers), Zhang et al. (2023) and other researchers (Young & Monroe, 2019) confirmed that perceived machine agency does have a role in explaining people's harsher moral judgments toward machine drivers. However, the LLMs did not mention any perception of machine driver agency when giving reasons for their moral judgments (see Tables S9–S14 in the SM).

We used GPT-4o (gpt-4o-2024-05-13, accessed via API) as a text analysis tool (see Oliveira et al., 2024) to extract, summarize, and compare the reasons given by LLMs in the human and machine driver conditions (the identity of the LLMs was masked before prompting GPT-4o). The automatic text analysis showed similarities and differences in the reasons provided by the LLMs in these two conditions (see Tables S9–S14 in the SM). Notably, the differences in these reasons did not reflect any of the human reasons (e.g., negative affect, perceived agency of machine drivers) mentioned

previously (Hidalgo et al., 2021; Liu & Du, 2022; Zhang et al., 2023). However, the qualitative differences in the reasons (e.g., while attributing causal responsibility to the human/machine drivers in Study 3, GPT-4 mentioned the human driver's inability to handle the situations safely and the failure of the automated system to handle the situation safely; see Table S14) were less informative about the root causes of the LLMs' bias against machine drivers. More probing efforts are needed to understand the moral "reasoning" applied or attributed (possibly in a post-hoc manner) by the LLMs.

Previous research has also noted that although LLMs can output causal and moral judgments that align with those of human participants, they seem to weigh the factors influencing their judgments differently than human participants (Nie et al., 2023) or display irrationally in ways that humans do not (Macmillan-Scott & Musolesi, 2024). Humans rely on both emotions and reason for moral judgments (Monin et al., 2007), whereas LLMs or other non-human agents are not expected to have human-like emotions driving their moral judgments. Thus, the mechanism or process by which LLMs arrive at moral responses might differ from those used by humans, despite similar outputs in terms of attributions of blame and harm. If so, this points to a fundamental insight for how we should understand LLMs or broader AI: their human-like behaviors and responses may result from different underlying processes or mechanisms.

In addition, we noted a difference between the two versions of GPTs: GPT-4, rather than GPT-3.5, consistently exhibited a human-like negativity bias against machine drivers. This interesting difference has practical significance. Previous investigations (Berent & Sansiveri, 2024; Zhao et al., 2023) also showed that more powerful GPT models tend to exhibit *stronger* biases, such as stereotype bias (Zhao et al., 2023). While developing LLMs, we naturally expect that LLMs can reflect human values and judgments as much as possible, and we are eager to reduce human-LLM misalignment. While human-LLM misalignment is risky, misguided alignment may also create even greater risks (Almeida et al., 2024). In reality, it is natural for people to turn to more powerful language models for what they anticipate will be "better" moral advice; however, according to our results, what they will get could be more biased output from these models in certain cases.

In summary, we tasked two LLMs with judging other machines (i.e., automated or driverless cars) when their actions result in harm. GPT-4 (but not GPT-3.5) consistently demonstrated human-like biased moral judgments against the other machines. The human-like bias may not simply reflect previously embedded human biases in training materials. Instead, it may derive from a distinct process or mechanism (perhaps functionally analogous to a kind of "moral reasoning") employed by the LLMs. Our work provides insights into understanding LLM-human (mis)alignment in moral judgments, contributing to the emerging fields of machine behavior and psychology (Bonnefon et al., 2024; Hagendorff et al., 2024; Rahwan et al., 2019; Yam et al., 2025).

Our work has several limitations. First, solely considering OpenAI's LLMs may limit generalizability. Focusing on automated and driverless cars as implicit moral machines raises uncertainty about GPT-4's bias generalizing to other machines or situations. Second, the evolving nature of LLMs also raises concerns about our findings' reproducibility over time (see the different results obtained from GPT-3.5 and GPT-4). Similar to our work in the moral domain, research (e.g., Chen et al., 2024) comparing GPT-3.5 and GPT-4 in non-moral domains also showed that LLM behavior can change over a relatively short period. Third, the LLM responses, as well as human responses, might be unstable when being asked in different ways (Röttger et al., 2024); thus, future work should consider diverse ways to evaluate the prevalence of LLMs' aversion against machines. Finally, we only considered bias in moral domains; future work should examine potential human-LLM differences in biases in non-moral domains (e.g., MacNeil et al., 2024; Sun et al., 2025).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Materials, data, code, results, and prompts are accessible at https://osf.io/5j9ux/?view_only=ab502bd6412744739ca2080b3ed5313c. In each study, we provide comprehensive information on conditions, data exclusions, and measures in this paper or its supplementary material.

Acknowledgements

Thank you to Guilherme Almeida and Jean-François Bonnefon for helpful feedback on an earlier draft and to Shuaiqi Chen and Wenting Tang for data collection. J.S. is supported by ANTITHESES Wellcome (Grant Number: 226801/Z/22/Z). This research is also supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-GV-2023-012), and supported by the Zhejiang University Qiushi Feiying Program and Zhejiang University Qiushi Xinxing Program.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, *120*(44), e2313790120.

Allen, J. W., Earp, B. D., Koplin, J., & Wilkinson, D. (2024). Consent-GPT: Is it ethical to delegate procedural consent to conversational AI? *Journal of Medical Ethics*, *50*(2), 77–83.

Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2024). Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, *333*, 104145.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., . . . Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*, 59–64.

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*. Virtual Event, Canada.

Berent, I., & Sansiveri, A. (2024). Davinci the dualist: The mind–body divide in large language models and in human learners. *Open Mind*, *8*, 84–101.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557.

Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology*, *75*(1), 653–675.

Chalmers, D. J. (2023). Could a large language model be conscious? *arXiv preprint*, arXiv:2303.07103.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., . . . Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 39.

Chemero, A. (2023). LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, *7*, 1828–1829.

Chen, L., Zaharia, M., & Zou, J. (2024). How Is ChatGPT's behavior changing over time?. *Harvard Data Science Review*, *6*(2).

Demaree-Cotton, J., Earp, B. D., & Savulescu, J. (2022). How to use AI ethically for ethical decision-making. *The American Journal of Bioethics*, *22*(7), 1–3.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600.

Earp, B. D., Sebastian, P. M., Jemima, A., Sabine, S., Vynn, S., Karin, J., . . . and Savulescu, J. (2024). A personalized patient preference predictor for substituted judgments in healthcare: Technically feasible and ethically desirable. *The American Journal of Bioethics*, *24*(7), 13–26.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17.

Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *iScience*, *24*(4), 102252.

Fujimoto, S., & Takemoto, K. (2023). Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence*, *6*, 1232003.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109.

Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., . . . Schulz, E. (2024). Machine psychology. *arXiv preprint*, arXiv:2303.13988.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*, 833–838.

Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2024). AI language models cannot replace human research participants. *AI & Society*, *39*, 2603–2605.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Virtual Conference.

Hidalgo, C. A., Orghiain, D., Canals, J. A., de Almeida, F., & Martin, N. (2021). *How Humans Judge Machines*. Cambridge, MA: The MIT Press.

Hong, J.-W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768–1774.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER*, 31122.

Liu, P., Chu, Y., Zhai, S., Zhang, T., & Awad, E. (2025). Morality on the road: Should machine drivers be more utilitarian than human drivers? *Cognition*, *254*, 106011.

Liu, P., & Du, Y. (2022). Blame attribution asymmetry in human–automation cooperation. *Risk Analysis*, *42*(8), 1769–1783.

Liu, P., Du, Y., & Xu, Z. (2019). Machines versus humans: People's biased responses to traffic accidents involving self-driving vehicles. *Accident Analysis & Prevention*, *125*, 232–240.

Macmillan-Scott, O., & Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, *11*(6), 240255.

MacNeil, S., Rogalska, M., Leinonen, J., Denny, P., Hellas, A., & Crosland, X. (2024). Synthetic students: A comparative study of bug distribution between large language models and computing students. In *Proceedings of the 2024 on ACM Virtual Global Computing Education Conference V. 1*. Virtual Event, NC.

McShane, B. B., & Böckenholt, U. (2022). Meta-analysis of studies with multiple contrasts and differences in measurement scales. *Journal of Consumer Psychology*, *32*(1), 23–40.

Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology*, *11*(2), 99–111.

Nie, A., Zhang, Y., Amdekar, A., Piech, C., Hashimoto, T., & Gerstenberg, T. (2023). MoCa: Measuring human-language model alignment on causal and moral judgment tasks. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. New Orleans, LA.

Oliveira, M., Brands, J., Mashudi, J., Liefooghe, B., & Hortensius, R. (2024). Perceptions of artificial intelligence system's aptitude to judge morality and competence amidst the rise of Chatbots. *Cognitive Research: Principles and Implications*, *9*, 47.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., . . . Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint*, arXiv:2402.16786.

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, *12*(3), 148.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, HA.

Scherrer, N., Shi, C., Feder, A., & Blei, D. (2023). Evaluating the moral beliefs encoded in LLMs. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. New Orleans, LA.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(4), 147–151.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258–268.

Shrawgi, H., Rath, P., Singhal, T., Dandapat, S., Graham, Y., & Purver, M. (2024). Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. St. Julian's, Malta.

Stella, M., Hills, T. T., & Kenett, Y. N. (2023). Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, *120*(43), e2312911120.

Stojilović, D., Franklin, M., Malle, B. F., Fernandez-Basso, C., Awad, E., & Lagnado, D. (2024). Are autonomous vehicles blamed differently? In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Rotterdam, the Netherlands.

Sun, F., Li, N., Wang, K., & Goette, L. (2025). Large Language Models are overconfident and amplify human bias. *arXiv preprint*, arXiv:2505.02151.

Takemoto, K. (2024). The moral machine experiment on large language models. *Royal Society Open Science*, *11*(2), 231393.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, *47*(7), e13309.

Vandersluis, R., & Savulescu, J. (2024). The selective deployment of AI in healthcare. *Bioethics*, *38*(5), 391–400.

Wan, Y., & Chang, K.-W. (2024). White men lead, black women help: Uncovering gender, racial, and intersectional bias in language agency. *arXiv preprint*, arXiv:2404.10508.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Vancouver, BC, Canada.

Yam, K. C., Eng, A., & Gray, K. (2025). Machine replacement: A mind-role fit perspective. *Annual Review of Organizational Psychology and Organizational Behavior*, *12*, 239–267.

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, 103870.

Zhang, J., Conway, J., & Hidalgo, C. A. (2023). Why people judge humans differently from machines: The role of perceived agency and experience. In *14th IEEE International Conference on Cognitive Infocommunications*. Budapest, Hungary.

Zhao, Y., Wang, B., Zhao, D., Huang, K., Wang, Y., He, R., & Hou, Y. (2023). Mind vs. mouth: On measuring re-judge inconsistency of social bias in large language models. *arXiv preprint*, arXiv:2308.12578.

Yueying Chu is a PhD candidate at the Center for Psychological Sciences and the Department of Psychology and Behavioral Sciences, Zhejiang University. She is a visiting PhD student at the Centre for Biomedical Ethics, National University of Singapore, since 2025. Her research focuses on machine psychology and moral psychology.

Peng Liu is with the Center for Psychological Sciences, Zhejiang University, China. He currently focuses on interdisciplinary and multidisciplinary research on AI machines and explores their relationships and interactions with humans and society.

Julian Savulescu is an award-winning ethicist and moral philosopher. He is the Chen Su Lan Centennial Professor in Medical Ethics and Director of the Centre for Biomedical Ethics at the Yong Loo Lin School of Medicine, National University of Singapore. His research has been published in high-profile journals such as *Nature*.

Brian D. Earp is an Associate Professor of Biomedical Ethics at the National University of Singapore, with courtesy appointments in philosophy and psychology. Brian's work is cross-disciplinary, drawing on training in philosophy, cognitive science, experimental psychology, history and sociology of science and medicine, and ethics.