Moral judgments on human vs. robot agents:

The role of a humanitarian reason

Yinuo Mu (Graduate School of Informatics, Nagoya University, mu.yinuo.a2@s.mail.nagoya-u.ac.jp, Japan)
Minoru Karasawa (Future Value Creation Research Center, Nagoya University, mkarasawa91@gmail.com, Japan)

人間とロボットの行為者に対する道徳的判断
—人道的理由の役割—

穆 一諾（名古屋大学 大学院情報学研究科）
唐沢 穣（名古屋大学 価値創造教育研究センター）

要約

人工知能やロボットがこれまでにないほど自律的になるにつれ、それらに対する人々の認識を研究する重要性が高まっている。本研究では、行為者の最初の判断に関連する文脈情報が、意図的でない結果に対する評価に影響を与えるかどうかを調べた。運転手（人間またはロボット）が人道的理由を持って（「理由あり」条件）交通規則を違反し、結果として歩行者をはねてしまうというシナリオを用いた。一方、「理由なし」条件では、運転手が赤信号を見落とすという過失により交通事故を起こした。日本の参加者は、2（理由の有無）×2（運転手の種類）の被験者間デザインのいずれかの条件にランダムに割り当てられ、運転手の非難と心理状態（意図性）について判断を求められた。結果として、理由なし条件では、人間運転手の非難度と意図性が高いことが示された。また、人道的理由という文脈情報が人間とロボットの対象に非対称な効果をもたらし、判断の差異を減少させた。具体的には、理由の存在により、ロボットに対する意図性の推論が向上された一方で、人間に対する非難評価が緩和された。本研究で使用された人道的理由（病人を助けること）は、人間にとっては免責される正当な理由である一方で、ロボットにとってはそうではないことを示された。また、ロボットの心的状態の推論は、人間に比べて文脈に大きく依存することが示唆された。

Key words
moral judgment, mental state attribution, machine morality, autonomous vehicles, human-robot interaction

1. Introduction

The question of whether a robot might be excused for a humanitarian reason when violating a rule is an important topic to explore. For instance, a robot is driving a car taking a severely sick woman to the hospital when it comes across a red light. At this moment, the robot is facing a dilemma: either to stop the car and thereby risk the sick woman's life, or run through it and thereby risk causing an accident. How would people judge the robot driver's decision in such a moral dilemma?

With artificial intelligence (AI) and robotic technologies being increasingly integrated into everyday life, their unprecedented impacts on human society have demanded greater attention. It is increasingly likely for a robot to face a dilemma such as the one mentioned above. Previous research has shown that as robots exhibit higher levels of autonomy, people tend to regard them as moral agents and hold them morally accountable for their actions (Kneer & Stuart, 2021; Sullins, 2011). However, the psychological mechanisms underlying such moral judgments remain largely unclear. In this study, we investigated how people attribute moral responsibility to autonomous machines, focusing on the roles of a humanitarian reason for violating a rule in a moral dilemma. We compared blame attribution toward human and robot targets for an unintended negative outcome and argued that reasons related to initial actions, serving as context information, prompt people to infer the mental states of an agent and to make blame judgments accordingly.

1.1 Blame judgment for robots compared to humans

Evidence on the blame judgment toward artificial agents compared to humans yielded mixed results. On the one hand, several studies found that people are prone to make harsher judgments for robot targets than for humans when they make the same decisions (e.g., Franklin et al., 2021; Malle et al., 2025; A. D. Young & Monroe, 2019). Scenarios used in these studies typically included explicit external causes such as avoiding a jaywalking pedestrian (Hong et al., 2020) or reasons such as saving more people (Malle et al., 2025; Young & Monroe, 2019) as context for the agents' mistakes. It is suggested that these external causes and reasons considerably relieved the human driver's moral blame. These studies predominantly came from the field of moral psychology, and some from the literature on human-robot interaction (HRI).

In contrast, studies predominantly from the literature on autonomous vehicles (AVs) suggest that human agents are assigned more blame and responsibility than artificial agents, such as autonomous cars (e.g., Awad et al., 2020; Beckers et al., 2022; Liu et al., 2021; Wotton et al., 2022). These studies are absent of specific external causes. Instead, the scenarios often involve accidents that occur because of inattentive human drivers failing to take control of semi-AVs in emergencies. For example, Awad and colleagues (2020) used a hypothesized double-driver car to mimic various levels of self-driving vehicles (US EPA, 2019). Their scenarios did not explicitly indicate any external cause or reason, nor did they clarify whether the drivers failed to avoid hitting the pedestrian due to an intentional decision. Despite the ambiguity, the patterns of blame attributions remain the same. Without an external cause or reason, the human drivers were condemned for not paying enough attention, and the judgments of blame were boosted.

Derived from this double-driver template, our previous research (Mu & Karasawa, 2024) extended these results to a situation similar to the dilemma mentioned in the beginning, where drivers had to choose whether to try to save a severely sick passenger by rushing to the hospital or to follow the traffic rule and avoid potential causing an accident. Results consistently indicated that when they chose to violate the rule and ended up hitting a pedestrian, the robot drivers were attributed less blame and the mental state of intending to run the red light than the humans who made the same mistake, which was referred to as the "humanness effect." In addition, this research demonstrated that across studies, the discrepancies in blame judgment between human and robot targets were accounted for by the attributed mental states. To the extent that the human drivers were inferred to have had more intentionality for the initial action (i.e., running through the red light), they were assigned more blame than their robot counterparts. Just like the perceived mental state of intentionality is a significant basis for judging humans (Monroe & Malle, 2017), it predicted the blame judgment toward robot agents as well.

Despite these findings, the double-driver settings might have been confounding with the humanness effect or the effect of the presence of the dilemma. Specifically, a human sub driver (who was monitoring the situation and ready to take control when it was necessary) could share more blame, leading to the corresponding main driver (who was primarily operating the car) being excused to a greater extent when the main driver was a robot. In other words, a robot main driver might be blamed less, either because its human sub driver shared more responsibility or because the robot itself was excused. The presence of sub drivers could also overshadow the effect of the moral dilemma which failed to show consistency across drivers, in that participants might have been reluctant to attribute the accident to external excuses when there were two responsible agents. To investigate the robustness of the interested effects, in the current study, we simplified the experiment by using only one driver and attempted to further emphasize the context of a moral dilemma. In line with our previous findings, we expected the "humanness effect" as an overall tendency, that is, human drivers would be judged as more blameworthy than robots (Hypothesis 1).

1.2 Attribution of mental state (i.e., intentionality) and the process of blame judgment for human vs robot agents

People tend to morally evaluate others' behaviors by associating these behaviors with the mental states behind them, such as beliefs, intentions, and emotions (Malle et al., 2014; Quesque et al., 2024). This study focuses exclusively on one particular construct: judging whether an observed action was deliberate or purposeful (i.e., the intentionality of the initial decision). Research on whether the relation between this attribution of intentionality and blame judgment extends to robot targets remains limited.

It is conceivable that nonhuman targets, such as robots, can be seen as having less mental agency as compared with humans. Such intuition is empirically supported: People have demonstrated being prone to at least partly denying nonhuman targets' mental capacities both consciously (Gray et al., 2007) and subconsciously (Li et al., 2022). Such a reduction in perceived mental capacities may result in a general tendency in which mental state attributions for robot targets are impaired. Indeed, in our previous research (Mu & Karasawa, 2024), participants attributed less intentionality to robot drivers than to humans. Hence, consistent with these findings, we expected that robots would be perceived as having less intentionality than humans (Hypothesis 2a).

In addition, we argue that the perception of intentionality accounts for the differences in blame judgments between humans and robots. A rich body of research has shown that intentionality attribution plays a critical role in the judgment of blame. It has been repeatedly confirmed that compared to accidental mistakes, intentional harm can induce strong condemnations and harsh judgments (Ames & Fiske, 2013; Lagnado & Channon, 2008; Monroe & Malle, 2017). This relationship can extend to robot targets also. Voiklis et al. (2016) used open-ended questions and asked participants why they had assigned blame to a human or robot target. They found that participants provided similar reasons for their blame judgments towards both human and robot targets, of which 59 % referred to mental agency. It is reasonable to speculate that the underlying mechanisms of blaming a human vs. robot agent are similar.

In our previous research (Mu & Karasawa, 2024) we found a positive correlation between attributions of mental state and blame that were consistent across the studies. Human drivers were assigned greater blame and responsibility for the accident, to the extent that they were attributed more mental states to violate the traffic rules than were the robots. Hence, in the current study, we hypothesized that the perception of intentionality would positively predict blame attributed to drivers (Hypothesis 2b). Together with Hypothesis 2a, we expected that the perceived intentionality would be a mediator in this process (Hypothesis 2, the mediation hypothesis.)

1.3 The impact of a humanitarian reason on the perception of intentionality

In the scenario mentioned at the beginning, the driver violates the traffic rule because they want to take an ill passenger to the hospital. Although the previous theory on blame judgment believes that the reasons for actors' behaviors will be considered only when they intend for negative outcomes (Malle et al., 2014), this article argues that, even when the consequences are unintended, the reasons related to the agent's initial behavior play a significant role in this process. In our previous research (Mu & Karasawa, 2024), we explored this question and found preliminary results that, compared to the negligent situation (i.e., inattentive drivers who merely failed to see the red light), drivers who had a reason were rated higher on both the intentionality of running the red light and the judgment of blame. In other words, the reasons associated with initial behaviors (i.e., violating the traffic rule) were taken into account when forming the attribution of mental state and, therefore, blame judgment, notwithstanding that the drivers had no intention to run over the pedestrian.

The current study further argues that the humanitarian reason in question may have different impacts on the perception of intentionality for humans versus robots. It is ubiquitous that people make sense of others' behaviors by inferring their mental states. This process is spontaneous and rapid: Within mere hundreds of milliseconds, people ascertain the mental state of an actor implied by a description of the actor's actions (Kruse & Degner, 2021; Van Overwalle et al., 2012; Young & Saxe, 2009). Meanwhile, people constantly infer others' actions as being intentional, even when the action is described as ambiguous, which is referred to as "intentional bias" (Rosset, 2008). People may proactively and effortlessly infer a human target's mental state and high intentionality, even with minimum information.

In contrast, intentionality bias might not happen for robot targets; rather, the inference of mental states for robots may heavily rely on context. Without additional cues, people may find it difficult to ascribe as many mental capacities to robots as to humans (Gray et al., 2007) or to imagine robots as having certain kinds of mental states, such as desire (de Graaf & Malle, 2019). Moreover, harmful outcomes like hitting a pedestrian, as in the current study, could decrease the mental states attributed especially to the robots (Stuart & Kneer, 2021), which may further lower the baseline for perceiving mental states in robots. However, with sufficient situational cues, a robot that can deal with a complex moral dilemma may exhibit a high level of sophistication and autonomy which can induce perceivers' ascription of mental capacities to the robot, thus attributing more mental states (Dawtry & Callan, 2024). In other words, the attribution of mental states toward robots could be more sensitive to context information (e.g., helping a sick passenger) compared to that toward humans; presenting a reason might reduce the difference between humans and robots. Therefore, in addition to the main effect of the type of agent on the perception of intentionality, we further hypothesized that with a humanitarian reason presented (i.e., helping an ill person), the perception of intentionality for robots would increase as compared with no reason being provided (i.e., negligence), whereas the perception for humans would hold constant (Hypothesis 3). The expected mediation would be mitigated with the humanitarian reason presented, compared to no reason presented. In other words, in addition to the mediation hypothesis, the indirect effect of driver type (human versus robot) on blame judgment through perceived intentionality would be moderated by the presence of reason, such that the indirect effect of driver type on blame judgment through perceived intentionality would be weaker for the drivers who intended to help the sick person than inattentive drivers (Hypothesis 4, the moderated mediation hypothesis).

1.4 Overview

In the current research, we aimed to emphasize the impacts of presenting a humanitarian reason on the attribution of mental state (i.e., intentionality) and blame toward human versus robot drivers. We directly compared a human with a robot driver, each of whom was solely controlling the car in a specific driving scenario. In the reason-present condition, the drivers were facing a moral dilemma (i.e., choosing either to stop at an intersection with a red light and thereby risking the life of the sick woman passenger, or to drive through the red light and take the sick woman to the hospital as quickly as possible and thereby risking causing an accident) and choosing to rush through the red light. In the reason-absent condition, the driver ran through the red light because of simply being negligent and ignoring the traffic signal. Participants' perceptions of intentionality and judgment of blame for drivers were assessed.

2. Method

2.1 Participants

Drawing on previous research, for the current study, we aimed to recruit 50 participants to evaluate each of the four scenarios. A total of 341 Japanese participants were recruited in Japan via the crowdsourcing platform, CrossMarketing. Of these, 73 were excluded for failing to finish the survey or disagreeing with their data being used for analysis, 62 were excluded for failure on manipulation check, and five were excluded because of a duration longer than 6,000 seconds or less than 210 seconds, leaving 201 participants (103 females, $M_{\text{age}} = 43.86$, $SD = 12.60$) subjected to the following analyses.

2.2 Experimental design

This experiment employed a 2 (humanness of the driver: human or robot) × 2 (reason: present or absent) between-subject factorial design based on random assignment.

2.3 Materials and procedure

Upon informed consent and indicating their demographic information (age, gender, and nationality), participants were presented with a picture of a driver, either a human or a robot, standing by car, and read a description of the accident that was written below the picture on the same page. In the reason-present condition, it was written that the driver was taking a severely sick woman to the hospital, whereas in the reason-absent condition, there was no such information provided. The description of the accident either stated that upon entering an intersection with a red light, the driver rushed through it either to get to the hospital as quickly as possible because otherwise the sick lady could die (in the reason-present condition) or due to negligence in seeing the red light but failing to stop (in the reason-absent condition). At the end of all scenarios, the driver who ignored the red light hit and badly injured a pedestrian who was crossing the street. In all conditions, when entering the intersection, the drivers saw the red light but failed to see the pedestrian who was about to cross the road, such that the pedestrian being injured was accidental. Participants had to stay on this page for more than 30 seconds before they could move on to the next page.

After reading the vignettes, participants judged the driver's blameworthiness (assessed with three questions). They were asked to indicate the degree to which they agreed that "The driver/robot should be blamed for this accident." "The driver/robot is responsible for causing the accident." and "The driver/robot is the cause for the accident." (Cronbach's α = .84). They also judged the driver' intentionality of violating the traffic rule (assessed with four questions) by indicating the degree to which they agreed that "The driver/robot ignored the red light intentionally." "The driver/robot intended to ignore the red light." "The driver/robot intentionally violated the traffic rule." and "The driver/robot intended to violate traffic rule." (Cronbach's α = .93). Questions for blame judgment and perceived intentionality were presented in random order and assessed on seven-point Likert scales ("1" = "Extremely disagree," and "7" = "Extremely agree"). Finally, participants were asked to report whether they had a driver's license and, if so, for how many years and how often they drove, ranging from zero to seven days a week. (See Supplemental Materials at https://osf.io/98g7d/?view_only=4b9e9512273e46c6b523a7456facf002 for the details of measurements). R packages "psych" (Revelle, 2023), "car" (Fox & Weisberg, 2019), "effectsize" (Ben-Shachar et al., 2020), and PROCESS for R (Hayes, 2022) were used in R 4.3.1 (R Core Team, 2023).

3. Results

3.1 Blame judgment.

We conducted a 2 × 2 ANOVA on blame. The type of driver showed a significant main effect (F (1,197) = 8.26, p < .01, η_p^2 = .04), supporting Hypothesis 1, whereas reason-present and reason-absent did not (F (1,197) = .45, p = .50). Their interaction yielded significance, with F (1,197) = 6.00, p < .05, η_p^2 = .03. The left panel of Figure 1 demonstrates the increased blame assigned to human drivers (M_{human} = 6.22, SD = .90) as compared to robot drivers (M_{robot} = 5.50, SD = 1.46, p < .01) in the reason-absent condition but not in the reason-present condition (M_{human} = 5.55, SD = 1.03, vs. M_{robot} = 5.66, SD = 1.31, p = .63). With the reason-absent condition held as the baseline, presenting a reason mitigated the blame assigned to the human driver ($M_{\text{-absent}}$ – $M_{\text{-present}}$ = .67, p < .01), but not that to the robot driver ($M_{\text{-absent}}$ – $M_{\text{-present}}$ = –.06, p = .82).

3.2 Perception of Intentionality.

A two-way ANOVA on the perception of intentionality yielded a significant main effect of the type of driver (F (1,197) = 27.37, p < .001, η_p^2 = .12) a significant main effect of the presence of reason (F (1,197) = 27.01, p < .001, η_p^2 = .12), and a significant interaction (F (1,197) =18.65, p < .001, η_p^2 = .09). Consistent with Hypothesis 2a, human drivers (M = 4.61, SD = 1.71) were judged as being more intentional than robot drivers (M = 3.79, SD = 2.00). The right panel of Figure 1 demonstrates the greater perceived intentionality to human drivers (M = 4.78, SD = 1.91) than to robot drivers (M = 2.85, SD = 1.61, p < .001) in the reason-absent condition; whereas in the reason-present condition, human (M = 4.49, SD = 1.57) and robot drivers (M = 4.71,

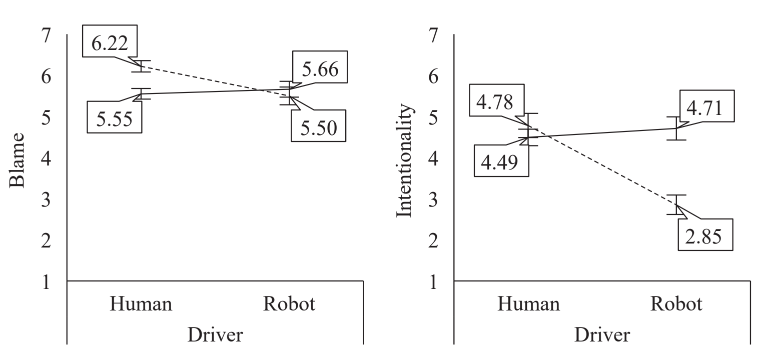

Figure 1: Results of analysis of variants (ANOVAs) on judgment of blame and intentionality.

Note: Error bars stand for standard errors.

SD = 1.93, *p* = .41) were judged as being equally intentional. Supporting Hypothesis 3, with the reason-absent condition held as the baseline, presenting a reason increased the intentionality perceived in the robot ($M_{\text{–absent}} – M_{\text{–present}}$ = –1.76, *p* < .0001), but not that of the human driver ($M_{\text{–absent}} – M_{\text{–present}}$ = .29, *p* = .40).

3.3 Moderated mediation analysis.

We conducted a moderated mediation analysis using Hayes' (2022) PROCESS for R Model 7 testing moderation (presence of the reason, with "reason-absent" dummy-coded as "0" and "reason-present" as "1") of the indirect path through perceived intentionality from the type of driver (with "robot" dummy-coded as "0" and "human" as "1") to blame judgment (bootstrapping *N* = 10,000). Hayes' index of moderated mediation = –.65, *SE* = .18, 95 %CI = [–1.02, –.32]. Further examination of the conditional indirect paths indicated that there was a significant indirect effect of humanness on blame judgment through perceived intentionality in the reason-absent condition, indirect effect = .58, *SE* = .15, 95 %CI = [.32, .89], whereas the effect was nonsignificant in the reason-present condition, indirect effect = –.07, *SE* = .10, 95 %CI = [–.27, .14]. In addition, upon accounting for the indirect effect, there was no significant direct effect (direct effect = .01, *SE* = .15, *p* = .97). Figure 2 indicates the breakdown of the conditional mediation model (also see Table 1 in the Supplemental Materials for the regression coefficients for each outcome variable). Together, these results partially support the expected moderated mediation. Supporting Hypothesis 4, the path through the type or driver to perceived intentionality was stronger in the reason-absent condition (consistent with Hypothesis 3), which resulted in a complete mediation, whereas such mediation diminished in the reason-present condition.

4. Discussion

This study documented people's responses to a humanitarian reason when judging humans' and robots' driving behaviors. Using a specific moral dilemma in which a humanitarian option conflicts with a traffic rule, the current study presents evidence of how this humanitarian reason when serving as context information related to the agents' initial decision affected the judgments on the decision's unintended consequences. While prior research suggested that the process of judging robot targets might involve similar normative expectations and mental models as judging humans (Komatsu, 2016; Voiklis et al., 2016), our results exhibited distinct processes between human and robot targets. When there's no sufficient information on the context of the agent's decision, people assign greater blame and mental states to humans than to robots. With this held as the baseline, learning the reason for the decision will induce increased mental states that are attributed to robots, whereas decreased blame judgment was allocated to humans. Eventually, humans and robots are judged as equally blameworthy and intentional when they are helping a sick person.

4.1 Distinct role of a humanitarian reason in humanness effect on blame judgment

Consistent with our previous findings (Mu & Karasawa, 2024), in one respect, the human drivers were blamed more than the robots, especially in negligence cases where the drivers ran the red light due to inattentiveness. This humanness effect took place only when there was no explicit reason for a rule violation, that is, the drivers were assumed to be inattentive (cf., Awad et al., 2020; Beckers et al., 2022). In another respect, the humanness effect decreased when presented with a humanitarian reason. The human and robot drivers were judged as being equally blameworthy when both had a severely sick passenger, due to the fact that blame judgment assigned to human drivers was attenuated, but not to those assigned to robot drivers. Helping a sick person seemed to be a good reason for human drivers to be excused, but not for robots to be excused.

This pattern contradicts some previous research in which harsher judgments were made for robot agents than for humans (Bigman & Gray, 2018; Young & Monroe, 2019), potentially in-

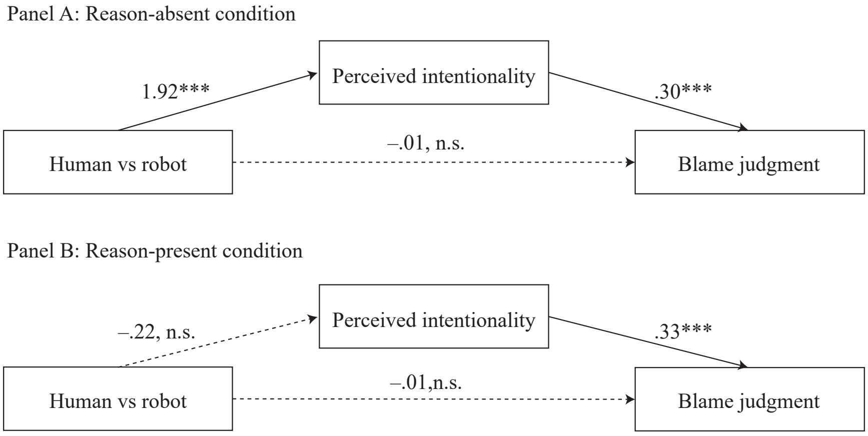

Figure 2: Results of moderated mediation analysis

Note: All path coefficients are unstandardized.

dicating boundary conditions. Although humans were somewhat excused, the reduction in moral judgment toward human targets in this study did not reach the level observed in prior research, where blame toward humans was even lower than that toward robots. It is possible that the effect of the humanitarian reason in the current study is not strong enough, as helping a sick person can only make a driver excusable to a limited extent. However, running the red light is considered a much worse violation that is not easily forgiven within Japanese culture where people expect others to follow the rules (Gelfand et al., 2011). As a result, even with humanitarian justification, the reduction in blame for human drivers appears moderate and never falls below the level of blame for robots. In contrast, Japanese people generally hold a positive attitude toward robots (Diana et al., 2023), which may account for the overall lower degree of blame for robots. Future research should explore the effects of different types of justifications interact with cultural preference for rule-compliance on moral judgments of humans versus robots.

Another potential reason why people did not attribute blame to robots as high an extent as had been demonstrated in previous research, is that people now tend to be more informed about the limitations of current robots and AI. It is possible that as AI applications have grown unprecedentedly quickly since the release of ChatGPT in 2022, people have become more familiar with AI (Inoue, 2024) and more tolerant of the mistakes they may make. Future research should investigate whether pre-experimental experience with AIs and robots has an impact on people's moral judgment toward robots.

The reason-present and reason-absent comparison has an important implication for the literature on blame judgment for accidental outcomes. For both scenarios, hitting a pedestrian was unintended, in other words, it was a side effect of running a red light. What was inconsistent between the conditions was the reason for violating the rule: When they violated the rule for a humanitarian reason, the human drivers were exonerated by a small but significant amount. This suggests that even though the transgressors did not intend to cause harm, observers still took into account the reason for the initial decision, contradicting the Path Model of Blame which predicts that a reason would be considered only when the outcome was intended (Malle et al., 2014).

4.2 Distinct role of a humanitarian reason in intentionality attribution toward human vs robot agents

People tend to infer high intentionality in humans across conditions, consistent with the well-documented phenomenon of "intentionality bias" (Rosset, 2008). It seems that this inference for humans is internally motivated: People proactively infer other's mental activities even with minimum contextual information. But when it comes to robots, this process seems to be context-dependent. Participants made judgments primarily based on the type of the driver only when there's no sufficient context information (i.e., reason-absent) as a cue for the driver's mental state. Individuals search for their experience with other people and/or the impressions drawn from the experimental materials. Laypersons tend to believe that a robot lacks some mental capacities (Gray et al., 2007), which might be a major obstacle to inferring mental states in a nonhuman target. It is easier to expect a greater mental state in human drivers compared to robots, and as a result, the human-robot difference emerges.

On the other hand, the attribution of mental states toward robots increases due to the presence of a humanitarian reason. It seems that people may use mental terms to explain robot behaviors, but they need an external prompt. Knowing that a robot is capable of dealing with complex situations such as prioritizing saving a human's life over the traffic rule, facilitates imagining that the robot driver in question has some thoughts that are very much like those of a human. As a result, the human-robot difference diminishes. Such a prompt fails to affect human drivers' assumed mental states, meaning that people might rely more heavily on the external context information for a robot agent than for a human.

One explanation for such an asymmetric effect is that robots are "humanized" due to the context information. We speculate that the perception of human and nonhuman objects such as machines and autonomous robots, lies along a continuum that indicates their levels of certain attributes (Bigman et al., 2023; Epley & Waytz, 2010) such as mental states. The presence of context information may shift the perceived position of a robot closer to that of a human along the continuum of attributed mental states. The more that people think that a robot and a human target are mentally alike, the more similar judgments people make toward the targets (Young & Monroe, 2019). When people believe that robots have a mind similar to that of humans, the differences in how people evaluate robots and humans may eventually disappear.

4.3 Limitations and future directions

Although participants did allocate blame to robot drivers, their motivations for this judgment remains unclear. From the social regulation perspective, blame judgment aims to correct others' mistakes so that they will conform to social norms in the future (Malle et al., 2014). As some researchers speculated, ordinary people may underestimate the easiness of reprograming robots and altering their behaviors (Bonnefon et al., 2023) and hence feel discouraged from regulating robots' behavior, which could be the reason why participants in the present study allocated less blame to the robot drivers. Future research may use different measures to probe the purpose of blaming the robots. If they aim for regulation, people's judgments may differ when they have a better understanding of how robots are programmed.

A key issue in this research concerns the pre-experiment condition of the participants' attitudes toward artificial agents. For instance, the tendency of anthropomorphism that is rooted in individuals' personal experiences can affect how much they perceive the existence of a mind in nonhuman targets (Epley et al., 2007). In addition, direct interaction with robots can promote positive attitudes toward them (Złotowski et al., 2015). Future research should include anthropomorphism tendencies and familiarity with robots as individual baselines.

It is also important to acknowledge that only one set of scenarios was used in this study. These results can likely extend to different situations of traffic violations conducted by autonomous vehicles (AVs), as well as other types of robots such as those used in healthcare settings. We agree that studies focused on other domains besides AVs should be emphasized as well (Bonnefon et al., 2023; Shank et al., 2019). However, we believe that the research on the mechanism underlying people's moral judgments toward AVs in various moral situations is far from thorough, especially in the current and future era where there are already self-driving cars on the road and people are not fully informed about the potential consequences.

5. Conclusion

This study revealed whether there is a reason presented as the context information related to an actor's initial behaviors, served as the basis of attribution of blame judgment through the inference of mental states, even when the negative outcome of accidentally hitting a pedestrian was unintended. In line with prior studies, we found that differentials in the attribution of mental states accounted for the humanness effect on blame judgment especially when the drivers ran the red light because of negligence. However, presenting the humanitarian reason for helping a sick woman passenger could lead to a decrease in the humanness effect, possibly due to the robot being perceived as being more like a human. Whereas inferring humans' mental states seems irrelevant to context information, people's inference of robots' intentionality can be facilitated by a humanitarian reason for which the humans were excused but the robots were not.

Acknowledgments

This work was supported by JST SPRING under Grant #JPMJSP2125; JSPS KAKENHI 22H01073 and a research grant from Canon Foundation. The authors report no conflict of interest.

References

Ames, D. L. & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science*, 24 (9), pp. 1755-1762.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, 4 (2), pp. 134-143.

Beckers, N., Siebert, L. C., Bruijnes, M., Jonker, C., & Abbink, D. (2022). Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid. *Scientific Reports*, 12 (1), 16193.

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5 (56), 2815.

Bigman, Y. E. & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181, pp. 21–34.

Bigman, Y. E., Surdel, N., & Ferguson, M. J. (2023). Trait attribution explains human-robot interactions. *Behavioral & Brain Sciences*, 46, pp. 1-65.

Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2023). The moral psychology of artificial intelligence. *Annual Review of Psychology*, 75 (1), pp. 653-675.

Dawtry, R. J. & Callan, M. J. (2024). Hazardous machinery: The assignment of agency and blame to robots versus nonautonomous machines. *Journal of Experimental Social Psychology*, 111, 104582.

de Graaf, M. M. A. & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. *Proceedings of 14th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 239-248.

Diana, F., Kawahara, M., Saccardi, I., Hortensius, R., Tanaka, A., & Kret, M. E. (2023). A cross-cultural comparison on implicit and explicit attitudes towards artificial agents. *International Journal of Social Robotics*, 15 (8), pp. 1439-1455.

Epley, N. & Waytz, A. (2010). Mind perception. In *Handbook of social psychology*, Vol. 1, 5th edition (pp. 498-541). John Wiley & Sons.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114 (4), pp. 864-886.

Fox, J. & Weisberg, S. (2019). Nonlinear regression, nonlinear least squares, and nonlinear mixed models in R. *Population*, 150, 200.

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

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M. S., Fischlmayr, I. C., Fischer, R., Fülöp, M., Georgas, J., Kashima, E. S., Kashima, Y., Kim, K., Lempereur, A., Marquez, P., Othman, R., Overlaet, B., Panagiotopoulou, P., Peltzer, K., Perez-Florizno, L. R., Ponomarenko, L., Realo, A., Schei, V., Schmitt, M., Smith, P. B., Soomro, N., Szabo, E., Taveesin, N., Toyama, M., Van de Vliert, E., Vohra, N., Ward, C., and Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332 (6033), pp. 1100-1104.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315 (5812), pp. 619-619.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach, 3rd edition*. The Guilford 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), pp. 1768-1774.

Inoue, Y. (2024). AI services growing in popularity among younger language learners in Japan. *The Japan Times* (Retrieved December 19, 2024 from https://www.japantimes.co.jp/news/2024/12/19/japan/chatgpt-english-lessons/).

Kneer, M. & Stuart, M. T. (2021). Playing the blame game with robots. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 407-411. Association for Computing Machinery.

Komatsu, T. (2016). Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 457-458. IEEE.

Kruse, F. & Degner, J. (2021). Spontaneous state inferences. *Journal of Personality and Social Psychology*, 121 (4), 774.

Lagnado, D. A. & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108 (3), pp. 754-770.

Li, Z., Terfurth, L., Woller, J. P., & Wiese, E. (2022). Mind the machines: Applying implicit measures of mind perception to social robotics. *2022 17th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 236-245. IEEE.

Liu, P., Du, M., & Li, T. (2021). Psychological consequences of legal responsibility misattribution associated with automated vehicles. *Ethics and Information Technology*, 23 (4), pp. 763-776.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, 25 (2), pp. 147-186.

Malle, B. F., Scheutz, M., Cusimano, C., Voiklis, J., Komatsu, T., Thapa, S., & Aladia, S. (2025). People's judgments of humans and robots in a classic moral dilemma. *Cognition*, 254, 105958.

Monroe, A. E. & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, 146 (1), pp. 123-133.

Mu, Y. & Karasawa, M. (2024). Blame attribution and intentionality perception of human versus robot drivers: Implications for judgments about autonomous vehicles in moral dilemma contexts. *Cogent Psychology*, 11 (1), 2384298.

Quesque, F., Apperly, I., Baillargeon, R., Baron-Cohen, S., Becchio, C., Bekkering, H., Bernstein, D., Bertoux, M., Bird, G., Bukowski, H., Burgmer, P., Carruthers, P., Catmur, C., Dziobek, I., Epley, N., Erle, T. M., Frith, C., Frith, U., Galang, C. M., Gallese, V., Grynberg, D., Happé, F., Hirai, M., Hodges, S. D., Kanske, P., Kret, M., Lamm, C., Nandrino, J. L., Obhi, S., Olderbak, S., Perner, J., Rossetti, Y., Schneider, D., Schurz, M., Schuwerk, T., Sebanz, N., Shamay-Tsoory, S., Silani, G., Spaulding, S., Todd, A. R., Westra, E., Zahavi, D., Bras, M. (2024). Defining key concepts for mental state attribution. *Communications Psychology*, 2 (1), pp. 1-5.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing (Retrieved January 22, 2025 from https://www.R-project.org).

Revelle, W. (2023). *How to use the psych package for regression and mediation analysis*.

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. *Cognition*, 108 (3), pp. 771-780.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing?: Moral attributions after individual and joint decisions. *Information, Communication & Society*, 22 (5), pp. 648-663.

Stuart, M. T. & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5, pp. 363:1-363:27.

Sullins, J. P. (2011). When is a robot a moral agent. *Machine Ethics*, 6 (2001), pp. 151-161.

US EPA, O. (2019). *Self-driving vehicles* [Other Policies and Guidance] (Retrieved May 17 from https://www.epa.gov/greenvehicles/self-driving-vehicles).

Van Overwalle, F., Van Duynslaeger, M., Coomans, D., & Timmermans, B. (2012). Spontaneous goal inferences are often inferred faster than spontaneous trait inferences. *Journal of Experimental Social Psychology*, 48 (1), pp. 13-18.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. robot agents. *Proceedings of 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 775-780.

Wotton, M. E. L., Bennett, J. M., Modesto, O., Challinor, K. L., & Prabhakharan, P. (2022). Attention all 'drivers': You could be to blame, no matter your behaviour or the level of vehicle automation. *Transportation Research Part F: Traffic Psychology and Behaviour*, 87, pp. 219-235.

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.

Young, L. & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, 21 (7), pp. 1396-1405.

Złotowski, J. A., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2015). Persistence of the uncanny valley: The influence of repeated interactions and a robot's attitude on its perception. *Frontiers in Psychology*, 6.

Received: April 8, 2025
Accepted: June 6, 2025
Published: December 25, 2025

https://doi.org/10.4189/shes.23.111