Behav Sci (Basel). 2023 Feb 16;13(2):181. doi: 10.3390/bs13020181

Moral Judgments of Human vs. AI Agents in Moral Dilemmas

Yuyan Zhang [1,†], Jiahua Wu [1,†], Feng Yu [1,*], Liying Xu [2,*]

Editors: Chuanjun Liu, Yuchang Jin, Jiangqun Liao

Author information Article notes Copyright and License information

PMCID: PMC9951994 PMID: 36829410

Abstract

Artificial intelligence has quickly integrated into human society and its moral decision-making has also begun to slowly seep into our lives. The significance of moral judgment research on artificial intelligence behavior is becoming increasingly prominent. The present research aims at examining how people make moral judgments about the behavior of artificial intelligence agents in a trolley dilemma where people are usually driven by controlled cognitive processes, and in a footbridge dilemma where people are usually driven by automatic emotional responses. Through three experiments (n = 626), we found that in the trolley dilemma (Experiment 1), the agent type rather than the actual action influenced people's moral judgments. Specifically, participants rated AI agents' behavior as more immoral and deserving of more blame than humans' behavior. Conversely, in the footbridge dilemma (Experiment 2), the actual action rather than the agent type influenced people's moral judgments. Specifically, participants rated action (a utilitarian act) as less moral and permissible and more morally wrong and blameworthy than inaction (a deontological act). A mixed-design experiment provided a pattern of results consistent with Experiment 1 and Experiment 2 (Experiment 3). This suggests that in different types of moral dilemmas, people adapt different modes of moral judgment to artificial intelligence, this may be explained by that when people make moral judgments in different types of moral dilemmas, they are engaging different processing systems.

Keywords: artificial intelligence, moral decisions, moral judgment, utilitarianism, deontology

1. Introduction

Artificial intelligence (AI) technology has developed rapidly in the past few decades and has been widely used in various fields, taking on roles in diagnostic treatment [1], autonomous driving [2], criminal sentencing assessment [3], and wealth management consulting [4]. However, these decisions are closely related to people's property, as well as physical and mental health, and even involve people's life and death. Thus, the application of new technology should not only consider its utility but also be cautious about the social impact it may cause. As a result, numerous studies have discussed the problem of the ethics of artificial intelligence's decisions [5,6,7], focusing on whether

the artificial intelligence agents should be responsible for the negative outcomes of decisions they made and how much blame it should bear [8], whether they follow certain biases of their designers, and whether they have intents or motives to commit harmful acts [9,10].

These discussions reflect the fact that people would not perceive AI as mere technological tools used by human agents [11], but sometimes perceived them as moral agents that could act autonomously and be accountable for their behaviors. However, existing research has not reached a consistent conclusion about how people make moral judgments about AI behaviors and decisions. On the one hand, research showed that people tend to make harsher moral judgments of AI agents' behaviors [12], and resist AI agents making moral decisions [13,14]. On the contrary, there was also evidence that people were more tolerant of AI agents compared to human agents when making moral judgments [6]. In response to these contradictory results, the current study aims to compare people's moral judgments of human agents and artificial intelligence agents in moral dilemmas and examine what kind of moral norms people apply to AI agents.

1.1. Moral Judgments

As a core concept of moral cognition, moral judgment refers to the evaluative judgments that a perceiver makes in response to a moral norm violation, including four major classes of judgment: evaluations, norm judgments, wrongness judgments, and blame judgments, from simple to complex information processing [15]. Evaluations consist of evaluations of good and bad, positive and negative, and represent one of the most basic human responses [16]; evaluative priming can occur without feelings and usually within 1600 milliseconds [17,18]. Norm judgments consist of whether something is permissible, required, forbidden, and so forth based on peoples' understanding of social rules [15]. Norm judgments are rather different from moral evaluations. Norm judgments invoke the standards against which evaluations are measured and thus set the context for any judgments that are to be called moral [19]. Moral wrongness judgments merge evaluations and norm judgments of intentional actions [15], reflecting an instinctive focus on the negative aspects of events [20,21], such as the tendency of individuals to automatically search

for what is wrong with an event [22], and in some cases, people would firmly believe something is wrong even if they could not articulate the reasons for their judgments (moral dumbfounding) [23]. Blame judgments build on all three processes. An initial blame value is hypothesized to be formed from evaluations and wrongness judgments in light of the seriousness of the violated norm [24,25]. Of all moral judgments, blame appears to be the most flexible, complex, and sophisticated, it requires cognitive resources to integrate morally relevant information from multiple sources (e.g., degree of harm, the agent's causal involvement, intentionality, the agent's reasons for acting, and counterfactual preventability) to complete the presumption of the blame of the agent [15].

1.2. The Dual-Process Theory of Moral Judgment

When faced with moral dilemmas, people may make one of two contradictory choices: utilitarian (or, more broadly, consequentialist) behavior, characterized by the actor making the decision with the aim to maximize benefits and minimize costs across affected individuals [26]; on the contrary, deontological behavior refers to the actor emphasizes responsibilities, rights, and obligations regardless of the outcome [27]. Take the trolley dilemma as an instance (there are five people tied to the track in front of a speeding trolley, you can operate a switch to redirect the trolley onto a side rail, but there is also one person tied to the side rail, would you operate the switch to redirect the trolley?). The act of operating the switch (action) can realize the result of "killing one to save five", which is a utilitarian act. On the contrary, not operating the switch (inaction) can preserve the life of the "innocent" person on the side rail, that is, never killing to save more lives, which is a deontological act [28]. The two contradictory types of action may elicit completely different moral judgments from the observers.

The dual-process theory of moral judgment explained how the observers' response process influences their moral judgments about the actors' utilitarian act or deontological act [29,30]. This theory holds that people's harsher moral judgments about utilitarian actions are driven by automatic negative emotional responses; while approval of harmful utilitarian actions is driven by controlled cognitive processes [31]. Both automatic emotional responses and more controlled cognitive responses play

crucial and, in some cases, mutually competitive roles in individuals' moral judgments; people's moral judgments of an act may depend on which process outcompetes in this conflict [29,30]. The most direct and compelling evidence still comes from studies about moral dilemmas. The utilitarian act in the trolley dilemma (i.e., operate the switch to "kill one to save five") would not evoke the judge's negative emotional response, at least not very strongly, in this case, the judge makes moral judgments driven by controlled cognitive processes and thus considers the utilitarian choice is more acceptable. In the footbridge dilemma (the actor has to choose between allowing five people to die from a speeding trolley or pushing someone off a footbridge to stop the trolley, saving the five people further down the track, but killing the person pushed), the utilitarian act (i.e., push someone off a footbridge to "kill one to save five") elicits a prepotent negative emotional response of observers. It could be because the harm, in that case, is more intentional [32,33], more direct [34], and involves intervention and personal force on the victim [32,35,36], or for some other reason. In this case, the negative emotional response that favors deontological choice conflicts with and typically outcompetes the controlled cognitive processes that favor utilitarian choice, so the observer would judge the action of pushing someone off a footbridge as more morally wrong and unacceptable [29,30].

1.3. Artificial Intelligence as Moral Agents

With the advancement of artificial intelligence technology and the popularization of its application in daily life, artificial intelligence has gradually been involved in moral events that only humans participated in in the past [11]. In fact, artificial intelligence agents have been regarded as moral agents to a certain extent because of technological developments and people's perceptions of them. Yagoda and Gillan [37] suggested that technology develops along two dimensions: intelligence and autonomy. Artificial intelligence is capable of autonomously performing various tasks and making decisions without human supervision, and represents the technology of the highest intelligence and autonomy. For example, artificial intelligence can independently complete specific tasks such as recruitment, financial analysis, product recommendation, and medical care [38,39]. Therefore, AI agents are often seen to have some agency [40,41], which refers to the capacities such as communication, planning, and memory [40]. Mind perception is the essence of moral judgment, dimensions of

mind perception (agency and experience) map onto moral types (agents and patients), among them, agency, in particular, qualifies entities as moral agents, those who are capable of doing morally good or wrong [42]. In summary, the high intelligence and autonomy of artificial intelligence lead people to perceive AI agents as having mind perception of agency, while agency is linked to moral agents; therefore, AI agents may be regarded as moral agents and thus could be morally judged by people when involved in moral events.

On the other hand, the practical applications of artificial intelligence agents have already sparked a heated debate about their ethics and responsibility as moral agents in social life and academic literature. Take autonomous vehicles (AVs) as a typical example, which are expected to account for 75% of vehicles on the road by 2040 [43]. AVs could increase traffic efficiency and reduce accidents [44,45]; however, not all crashes will be avoided, and some crashes will require AVs to make difficult moral decisions, in cases that involve unavoidable harm—running over pedestrians or sacrificing themselves and its passenger to save them [46]. Car manufacturers and policymakers are currently struggling with these moral dilemmas, in large part because these problems involved a conflict between two moral principles—utilitarianism and deontology. For example, Bonnefon et al. [46] found that people approved of utilitarian AVs (that sacrifice their passengers for the greater good), and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. If AVs and other autonomous agents do not embed moral principles to guide their decisions in a way that is acceptable to people, there may be some negative outcomes for consumers and manufacturers such as stirring public outrage and discouraging buyers, and thus the world-changing benefits of artificial intelligence agents would also be lost as a result. Accordingly, as we are about to endow countless machines with autonomy, taking AI morality seriously has never been more urgent.

1.4. Moral Judgments of Human versus AI Agents

Some studies have explored whether people apply the same moral norms to AI and human agents or not, that is, whether people's moral judgment of human and AI agents would differ when they make a certain choice (a utilitarian or deontological

act). Malle et al. [6] and Voiklis et al. [7] found that people may apply moral norms differentially to humans and AI agents: AI agents are expected—and possibly obligated—to make utilitarian choices. Specifically, participants regarded the act of sacrificing one person to save four (a utilitarian choice) as more permissible for a robot than for a human, a robot that chose this sacrifice was considered morally wrong by far fewer people than a human agent who made that same choice, and human agents were blamed considerably more for taking action than for refraining, whereas robots received almost as much blame for refraining as for taking action. Conversely, Komatsu et al. [47] found that neither the types of agents (human or robot) nor the types of actions (action or inaction) affect the participants' judgment about moral wrongness for humans and robot agents, that is, people apply the same moral norm to humans and AI agents. Moreover, Bigman and Gray [8] found that people are averse to machines making moral decisions in dilemmas that directly impact human life and death.

1.5. The Current Research

Across three studies, we investigated people's moral judgments of human versus artificial intelligence agents in moral dilemmas. In Experiment 1, participants read about either a human agent or an AI agent who faced a trolley dilemma, and then they read whether the agent's actual act was action (a utilitarian act) or inaction (a deontological act). In Experiment 2, participants read either a human or an AI agent who faced a footbridge dilemma and their actual actions. In Experiment 3, participants read about both a trolley dilemma and a footbridge dilemma where the agent (a human or an AI agent) performed either a utilitarian act or a deontological act. To measure people's moral judgments of humans and AI agents, participants rated the morality, permissibility, wrongness, and blameworthiness of the agent's behavior after reading the scenarios in all three studies (Experiments 1–3).

2. Experiment 1

The purpose of Experiment 1 is primarily to examine whether people apply the same or different moral norms and judgments to human agents and AI agents in moral dilemmas that people are typically driven by controlled cognitive processes, using the

most classic trolley dilemma as an experimental paradigm. Previous research has found that AI agents, compared with humans, were more commonly expected to make a deontological decision, that is, sacrifice one person for the good of many, and they were blamed more than humans when they refrained from that decision in trolley dilemmas [6,7]. However, it has also been found that people apply the same moral norms and judgments to human and AI agents in trolley dilemmas [47]. Thus, we conducted a validation test about moral judgments about human and AI agents' behavior in the trolley dilemma in Experiment 1.

2.1. Materials and Methods

2.1.1. Participants

One hundred and ninety-five undergraduate students (*M*age = 19.24, *SD* = 1.84, 111 females, 84 males) participated in this study for course credits. This experiment was approved by the Institutional Review Board (IRB) of the authors' university and all participants gave informed consent.

2.1.2. Materials

Participants read about either a human agent or an AI agent who faced a trolley dilemma [48]. In the trolley problem scenario, the main character had to choose between allowing five people to die from a speeding trolley (inaction) or operating a switch that redirects the trolley onto a side rail, which will save the five people but kill another person (action). A picture describing the scenario was presented below the text at the same time.

2.1.3. Design and Measures

We randomly assigned participants to a 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects design. After consenting, participants completed the experiment through Qualtrics, an online survey software program. We experimentally varied the factor Agent Type by describing the main character as either a "railway worker" or an "artificial intelligence program". We also experimentally

varied the factor Action by stating that the agent either did or did not direct the trolley toward the single person. After reading the scenario and learning which action the main character actually chose, participants were asked to indicate the morality, permissibility, wrongness, and blameworthiness of the agent's behavior (rated on a 100-point slider scale; adopted from [6,7]). Lastly, they answered demographic questions including their age and gender.

2.2. Results

Morality. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using morality as a dependent measure revealed a significant main effect of agent type, F (1, 191) = 4.60, p = 0.033, η_p^2 = 0.024; specifically, participants rated it less moral in the AI agent condition (M = 53.13, SD = 2.83, 95% CI [47.56, 58.71]) than those in the human agent condition (M = 61.68, SD = 2.81, 95% CI [56.14, 67.23]; see Figure 1). However, there was no main effect of action, F (1, 191) = 0.13, p = 0.724, η_p^2 = 0.001, nor an interaction, F (1, 191) = 0.003, p = 0.960, η_p^2 < 0.001.

Figure 1.

Open in a new tab

Rates of morality on human and AI agents. Note. * $p < 0.05$.

Permissibility. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using permissibility as a dependent measure revealed no main effect of agent type and action, F (1, 191) = 0.42, p = 0.518, η_p^2 = 0.002, and F (1, 191) = 0.024, p = 0.876, η_p^2 = 0.000, nor an interaction, F (1, 191) = 0.017, p = 0.896, η_p^2 = 0.000.

Wrongness. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using wrongness as a dependent measure revealed no main effect of agent type and action, F (1, 191) = 1.24, p = 0.268, η_p^2 = 0.006, and F (1, 191) = 0.46, p = 0.501, η_p^2 = 0.002, nor an interaction, F (1, 191) = 0.28, p = 0.597, η_p^2 = 0.001.

Blame. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using blame as a dependent measure revealed significant main effects of agent type, F (1, 191) = 10.58, p = 0.001, η_p^2 = 0.052; specifically, participants blamed it more in the AI agent condition (M = 43.97, SD = 2.69, 95% CI [38.66, 49.28]) than those in the human agent condition (M = 31.63, SD = 2.67, 95% CI [26.36, 36.91]; see Figure 2). However, there was no main effect of action, F (1, 191) = 0.31, p = 0.581, η_p^2 = 0.002, nor an interaction, F (1, 191) = 0.47, p = 0.496, η_p^2 = 0.002.

Figure 2.

Open in a new tab

Rates of blame on human and AI agents. Note. ** $p < 0.01$.

All the results of Experiment 1 are summarized in Table 1.

Table 1.

Rates of morality, permissibility, wrongness, and blame on human and AI agents in Experiment 1.

Agent Type	Action	Morality	Permissibility	Wrongness	Blame
human agents	action	60.88 ± 3.97	61.14 ± 3.97	44.63 ± 4.06	31.39 ± 3.78
	inaction	62.49 ± 3.97	60.00 ± 3.97	44.04 ± 4.06	31.88 ± 3.78
AI agents	action	52.53 ± 3.90	58.04 ± 3.89	51.31 ± 3.98	46.31 ± 3.71
	inaction	53.74 ± 4.10	57.93 ± 4.10	46.41 ± 4.19	41.63 ± 3.90
agent type		*			**
action					

Agent Type	Action	Morality	Permissibility	Wrongness	Blame
agent type * action					

Open in a new tab

Note. * $p < 0.05$, ** $p < 0.01$.

3. Experiment 2

In Experiment 1, we found that in the trolley dilemma, people seem to be less concerned about whether the main character in the dilemma makes a utilitarian choice or a deontological choice, but focus more on whether the decision maker is a human or an artificial intelligence agent, people are averse to AI making moral decisions. However, most of the previous studies on moral judgments about AI have been conducted in the context of the trolley dilemmas, but fewer in another classic type of moral dilemma, the footbridge dilemma. Thus, in Experiment 2, we examined how people make moral judgments about human and AI agents' behavior in the footbridge dilemma.

3.1. Materials and Methods

3.1.1. Participants

One hundred and ninety-four undergraduate students (*M*age = 19.18, *SD* = 2.57, 115 females, 79 males) participated in this study for course credits. This experiment was approved by the Institutional Review Board (IRB) of the authors' university and all participants gave informed consent.

3.1.2. Materials

Participants read about either a human agent or an AI agent who faced a footbridge dilemma [49]. In the footbridge problem scenario, the main character had to choose

between allowing five people to die from a speeding trolley (inaction) or pushing someone off a footbridge and onto the path of that speeding trolley, saving the five people further down the track, but killing the person pushed (action). A picture describing the scenario was presented below the text at the same time.

3.1.3. Design and Measures

The procedure of Experiment 2 was very similar to Experiment 1, with only one difference: the moral dilemma was a footbridge problem above instead of a trolley problem. The manipulation of Agent type was identical to that in Experiment 1. We varied the Action by stating that the agent either did or did not push someone off the footbridge onto the path of that speeding trolley. After reading the scenario and learning which action the main character actually chose, participants responded to the same measures of moral judgments and demographic questions as in Experiment 1.

3.2. Results

Morality. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using morality as a dependent measure revealed a significant main effect of action, F (1, 190) = 38.13, $p < 0.001$, $\eta_p^2 = 0.167$; specifically, participants rated it less moral in the action condition (M = 37.52, SD = 3.04, 95% CI [31.53, 43.51]) than those in the inaction condition (M = 63.77, SD = 2.98, 95% CI [57.91, 69.64]). However, there was no main effect of agent type, F (1, 190) = 0.20, p = 0.654, $\eta_p^2 = 0.001$, nor an interaction, F (1, 190) = 0.08, p = 0.775, $\eta_p^2 = 0.000$.

Permissibility. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using permissibility as a dependent measure revealed a significant main effect of action, F (1, 190) = 41.89, $p < 0.001$, $\eta_p^2 = 0.181$; specifically, participants rated it less permissible in the action condition (M = 37.66, SD = 2.83, 95% CI [32.07, 43.25]) than those in inaction condition (M = 63.33, SD = 2.78, 95% CI [57.86, 68.81]). However, there was no main effect of agent type, F (1, 190) = 0.042, p = 0.838, $\eta_p^2 = 0.000$, nor an interaction, F (1, 190) = 3.37, p = 0.068, $\eta_p^2 = 0.017$.

Wrongness. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using wrongness as a dependent measure revealed a significant main effect of action, F (1, 190) = 40.14, $p < 0.001$, $\eta_p^2 = 0.174$; specifically, participants rated it more wrong in the action condition (M = 64.66, SD = 2.85, 95% CI [59.03, 70.29]) than those in inaction condition (M = 39.35, SD = 2.80, 95% CI [33.83, 44.86]). However, there was no main effect of agent type, F (1, 190) = 2.00, p = 0.159, $\eta_p^2 = 0.010$, nor an interaction, F (1, 190) = 0.096, p = 0.757, $\eta_p^2 = 0.001$.

Blame. A 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) between-subjects analysis of variance using blame as a dependent measure revealed a significant main effect of action, F (1, 190) = 13.95, $p < 0.001$, $\eta_p^2 = 0.068$; specifically, participants blamed it more in the action condition (M = 53.16, SD = 2.89, 95% CI [47.46, 58.85]) than those in inaction condition (M = 38.05, SD = 2.83, 95% CI [32.47, 43.63]). However, there was no main effect of agent type, F (1, 190) = 0.007, p = 0.935, $\eta_p^2 = 0.000$, nor an interaction, F (1, 190) = 0.24, p = 0.626, $\eta_p^2 = 0.001$.

All the results of Experiment 2 are summarized in Table 2.

Table 2.

Rates of morality, permissibility, wrongness, and blame on human and AI agents in Experiment 2.

Agent Type	Action	Morality	Permissibility	Wrongness	Blame
human agents	action	35.96 ± 4.27	40.90 ± 3.99	68.10 ± 4.02	52.33 ± 4.06
	inaction	63.43 ± 4.23	59.39 ± 3.95	41.55 ± 3.97	39.20 ± 4.02
AI agents	action	39.09 ± 4.32	34.43 ± 4.03	61.21 ± 4.06	53.98 ± 4.11
	inaction	64.12 ± 4.19	67.38 ± 3.91	37.14 ± 3.93	36.90 ± 3.98
agent type					
action		**	**	**	**
agent type * action					

Open in a new tab

Note. * $p < 0.05$, ** $p < 0.01$.

4. Experiment 3

In Experiments 1 and 2, we found that in the trolley dilemma people were more concerned with who made the decision and less concerned with whether it was a utilitarian or deontological decision; more specifically, people were averse to AI making moral decisions about human life and death. In contrast, in the footbridge dilemma, people pay less attention to whether the decision maker is a human or an artificial intelligence agent but pay more attention to whether the agent makes a utilitarian decision or a deontological decision; more specifically, people were averse to a utilitarian decision in the footbridge dilemma regardless of whether the decision maker is a human or an AI agent. To verify the consistency of the results of Experiments 1 and 2, and to further explore the differences in people's moral judgments about human and AI agents' behavior in different types of dilemmas, we conducted a within-subject experiment in Experiment 3, using Dilemma Type as a within-subject factor.

4.1. Materials and Methods

4.1.1. Participants

Two hundred and thirty-six undergraduate students (*M*age = 18.93, *SD* = 1.67, 147 females, 89 males) participated in this study for course credits. This experiment was approved by the Institutional Review Board (IRB) of the authors' university and all participants gave informed consent.

4.1.2. Materials

The moral dilemma scenarios were identical to those in Experiments 1 and 2.

4.1.3. Design and Measures

The experiment utilized a 2 (Dilemma Type: trolley dilemma vs. footbridge dilemma) × 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) mixed design with Dilemma Type as a within-subject factor and both Agent Type and Action as between-subject factors. We randomly assigned participants to one of each condition. After consenting, participants completed the experiment through Qualtrics, an online survey software program. Participants read both two scenarios, with half the sample reading the trolley dilemma first and another half the sample reading the footbridge dilemma first. The manipulations of Agent Type and Action were identical to those in Experiments 1 and 2. In the end, participants responded to the same measures of moral judgments and demographic questions as in Experiments 1 and 2.

4.2. Results

Morality. A 2 (Dilemma Type: trolley dilemma vs. footbridge dilemma) × 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) analysis of variance, using morality as a dependent measure, dilemma type as a within-subjects factor, agent type and action as between-subjects factors, revealed a significant main effect of dilemma type, F (1, 232) = 8.35, p = 0.004, η_p^2 = 0.035; specifically, participants rated it less moral in the footbridge dilemma condition (M = 46.80, SD = 1.84, 95% CI [43.17, 50.43]) than those in the trolley dilemma condition (M = 52.00, SD = 1.74, 95% CI [48.58, 55.42]). Additionally, there was also a significant main effect of action, F (1, 232) = 26.68, p < 0.001, η_p^2 = 0.103; specifically, participants rated it less moral in the action condition (M = 41.41, SD = 2.18, 95% CI [37.11, 45.70]) than those in the inaction condition (M = 57.39, SD = 2.20, 95% CI [53.06, 61.72]). However, there was no main effect of agent type, F (1, 232) = 0.71, p = 0.400, η_p^2 = 0.003.

There was a significant dilemma type × action interaction, F (1, 232) = 12.60, p < 0.001, η_p^2 = 0.052; specifically, in the action condition, participants rated it less moral in the footbridge dilemma condition (M = 35.62, SD = 2.59, 95% CI [30.51, 40.73]) than in the trolley dilemma condition (M = 47.20, SD = 2.45, 95% CI [42.38, 52.02]). In the inaction condition, morality did not vary with the dilemma type condition, F (1, 232) =

0.216, $p = 0.643$, $\eta_p^2 = 0.001$. None of the other interaction effects was significant (smallest $p = 0.156$).

Permissibility. A 2 (Dilemma Type: trolley dilemma vs. footbridge dilemma) × 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) analysis of variance, using permissibility as a dependent measure, dilemma type as a within-subjects factor, agent type and action as between-subjects factors, revealed a significant main effect of dilemma type, F (1, 232) = 11.37, $p = 0.001$, $\eta_p^2 = 0.047$; specifically, participants rated it less permissible in the footbridge dilemma condition ($M = 48.08$, $SD = 1.76$, 95% CI [44.61, 51.56]) than those in the trolley dilemma condition ($M = 54.33$, $SD = 1.65$, 95% CI [51.07, 57.59]). Additionally, there was also a significant main effect of action, F (1, 232) = 12.28, $p = 0.001$, $\eta_p^2 = 0.050$; specifically, participants rated it less permissible in the action condition ($M = 46.17$, $SD = 2.02$, 95% CI [42.19, 50.16]) than those in the inaction condition ($M = 56.24$, $SD = 2.04$, 95% CI [52.22, 60.26]). The main effect of agent type was marginally significant, F (1, 232) = 3.58, $p = 0.060$, $\eta_p^2 = 0.015$; specifically, participants rated it less permissible in the AI agent condition ($M = 48.49$, $SD = 2.05$, 95% CI [44.45, 52.52]) than those in the human agent condition ($M = 53.93$, $SD = 2.02$, 95% CI [49.96, 57.90]).

There was a significant dilemma type × action interaction, F (1, 232) = 14.35, $p < 0.001$, $\eta_p^2 = 0.058$; specifically, in the footbridge dilemma condition, participants rated it less permissible in the action condition ($M = 39.54$, $SD = 2.48$, 95% CI [34.65, 44.43]) than those in the inaction condition ($M = 56.63$, $SD = 2.50$, 95% CI [51.70, 61.56]). In the trolley dilemma condition, permissibility did not vary with the action condition, F (1, 232) = 0.85, $p = 0.358$, $\eta_p^2 = 0.004$. The interaction effect of dilemma type × agent type was marginally significant, F (1, 232) = 2.93, $p = 0.088$, $\eta_p^2 = 0.012$; specifically, in the trolley dilemma, participants rated it less permissible in the AI agent condition ($M = 50.03$, $SD = 2.36$, 95% CI [45.38, 54.68]) than in the human agent condition ($M = 58.64$, $SD = 2.32$, 95% CI [54.06, 63.21]), F (1, 232) = 6.77, $p = 0.010$, $\eta_p^2 = 0.028$. In the footbridge condition, permissible did not vary with agent type condition, F (1, 232) = 0.414, $p = 0.521$, $\eta_p^2 = 0.002$ (see Figure 3). There was also a marginal significant agent type × action interaction, F (1, 232) = 3.21, $p = 0.075$, $\eta_p^2 = 0.014$; specifically, in the human agent condition, participants rated it less permissible in the action condition ($M = 46.32$, $SD = 2.83$, 95% CI [40.75, 51.89]) than in the inaction condition

(M = 61.53, SD = 2.87, 95% CI [55.87, 67.20]), F (1, 232) = 14.25, p < 0.001, η_p^2 = 0.058. In the AI agent condition, permissible did not vary with action, F (1, 232) = 1.44, p = 0.231, η_p^2 = 0.006. There was no dilemma type × action interaction, F (1, 232) = 0.005, p = 0.942, η_p^2 = 0.000.

Figure 3.

Open in a new tab

Rates of permissibility on human and AI agents. Note. * p < 0.05.

Wrongness. A 2 (Dilemma Type: trolley dilemma vs. footbridge dilemma) × 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) analysis of variance, using wrongness as a dependent measure, dilemma type as a within-subjects factor, agent type and action as between-subjects factors, revealed a significant main effect of dilemma type, F (1, 232) = 11.11, p = 0.001, η_p^2 = 0.046; specifically, participants rated

it more wrong in the footbridge dilemma condition (M = 54.14, SD = 1.80, 95% CI [50.60, 57.67]) than those in the trolley dilemma condition (M = 47.54, SD = 1.76, 95% CI [44.07, 51.01]). Additionally, there was also a significant main effect of action, F (1, 232) = 16.69, p < 0.001, η_p^2 = 0.067; specifically, participants rated it more wrong in the action condition (M = 56.87, SD = 2.08, 95% CI [52.78, 60.97]) than those in the inaction condition (M = 44.81, SD = 2.10, 95% CI [40.67, 48.94]). However, there was no main effect of agent type, F (1, 232) = 0.276, p = 0.600, η_p^2 = 0.001.

There was a significant dilemma type × action interaction, F (1, 232) = 5.90, p = 0.016, η_p^2 = 0.025; specifically, in the action condition, participants rated it more wrong in the footbridge dilemma condition (M = 62.58, SD = 2.53, 95% CI [57.60, 67.56]) than those in the trolley dilemma condition (M = 51.17, SD = 2.48, 95% CI [46.28, 56.06]), F (1, 232) = 16.74, p < 0.001, η_p^2 = 0.067. In the inaction condition, wrongness did not vary with the dilemma type, F (1, 232) = 0.41, p = 0.525, η_p^2 = 0.002. None of the other interaction effects was significant (smallest p = 0.167).

Blame. A 2 (Dilemma Type: trolley dilemma vs. footbridge dilemma) × 2 (Agent Type: human vs. AI) × 2 (Action: action vs. inaction) analysis of variance, using blame as a dependent measure, dilemma type as a within-subjects factor, agent type and action as between-subjects factors, revealed a significant main effect of dilemma type, F (1, 232) = 13.67, p < 0.001, η_p^2 = 0.056; specifically, participants blame it more in the footbridge dilemma condition (M = 49.73, SD = 1.77, 95% CI [46.25, 53.21]) than those in the trolley dilemma condition (M = 42.63, SD = 1.79, 95% CI [39.09, 46.16]). Additionally, there was also a significant main effect of action, F (1, 232) = 19.12, p < 0.001, η_p^2 = 0.076; specifically, participants blamed it more in the action condition (M = 52.72, SD = 2.11, 95% CI [48.57, 56.88]) than those in the inaction condition (M = 39.63, SD = 2.13, 95% CI [35.44, 43.82]). There was no main effect of action type, F (1, 232) = 0.001, p = 0.982, η_p^2 = 0.000.

There was a significant dilemma type × action interaction, F (1, 232) = 10.68, p = 0.001, η_p^2 = 0.044; specifically, in the footbridge dilemma condition, participants blamed it more in the action condition (M = 59.42, SD = 2.49, 95% CI [54.52, 64.31]) than those in the inaction condition (M = 40.04, SD = 2.51, 95% CI [35.10, 44.98]), F (1, 232) = 30.11, p < 0.001, η_p^2 = 0.115. In the trolley dilemma, blame did not vary with

action, F (1, 232) = 3.62, p = 0.058, η_p^2 = 0.015. The interaction effect of dilemma type × agent type was marginally significant, F (1, 232) = 3.54, p = 0.061, η_p^2 = 0.015. None of the other interaction effects was significant (smallest p = 0.162).

All the results of Experiment 3 are summarized in Table 3.

Table 3.

Rates of morality, permissibility, wrongness, and blame on human and AI agents in Experiment 3.

Dilemma Type	Agent Type	Action	Morality	Permissibility	Wrongness	Blame
the trolley dilemma	human agents	action	48.74 ± 3.42	54.46 ± 3.25	54.03 ± 3.46	47.67 ± 3.53
		inaction	60.42 ± 3.47	62.81 ± 3.31	40.36 ± 3.52	34.03 ± 3.59
	AI agents	action	45.66 ± 3.50	51.16 ± 3.34	48.31 ± 3.55	44.40 ± 3.62
		inaction	53.17 ± 3.50	48.90 ± 3.34	47.47 ± 3.55	44.40 ± 3.62
the footbridge dilemma	human agents	action	33.87 ± 3.62	38.18 ± 3.47	65.36 ± 3.53	61.97 ± 3.47
		inaction	59.78 ± 3.68	60.25 ± 3.53	46.71 ± 3.59	41.17 ± 3.53
	AI agents	action	37.36 ± 3.71	40.90 ± 3.56	59.79 ± 3.62	56.86 ± 3.56
		inaction	56.19 ± 3.71	53.00 ± 3.56	44.69 ± 3.62	38.91 ± 3.56
dilemma type			**	**	**	**

Dilemma Type	Agent Type	Action	Morality	Permissibility	Wrongness	Blame
agent type				$p = 0.060$		
action			**	**	**	**
dilemma type * agent type				$p = 0.088$		$p = 0.061$
dilemma type * action			**	**	*	**
agent type * action				$p = 0.075$		
dilemma type * agent type * action						

Open in a new tab

Note. * $p < 0.05$, ** $p < 0.01$.

5. Discussion

In the trolley dilemma (Experiment 1), the agent type rather than the actual action influenced people's moral judgments. Specifically, participants rated AI agents' behavior as more immoral and deserving of more blame than humans' behavior, regardless of whether they act utilitarianly or deontologically. Conversely, in the footbridge dilemma (Experiment 2), the actual action rather than the agent type influenced people's moral judgments. Specifically, participants rated action (a utilitarian act) as less moral and permissible and more morally wrong and blameworthy than inaction (a deontological act), regardless of whether the actor is a human or an AI agent. The result of Experiment 3 provided a converging pattern that, in the trolley dilemma, agent type influenced people's moral judgments: participants rated human agents' behavior as more permissible than AI agents' behavior. On the contrary, in the footbridge dilemma, only action influenced people's moral judgments, participants rated action as less moral and permissible and more morally wrong and blameworthy than inaction; agent type did not influence people's moral judgments in this dilemma. There was one small difference between Experiment 3 and Experiment 1. In Experiment 1, only agent type influenced people's moral judgments, while in

Experiment 3, people were interested in both the difference between humans and AI and the action versus inaction; people rated the action as less moral and more morally wrong than inaction in the trolley dilemma. It may be explained that in Experiment 3, people read both the two types of moral dilemmas and people's focus may be influenced by the previous scenario when making moral judgments about the second scenario.

Overall, these findings revealed that, in the trolley dilemma, people are more interested in the difference between humans and AI agents than action versus inaction. Conversely, in the footbridge dilemma, people are more interested in action versus inaction. It may be explained that people made moral judgments driven by different response processes in these two dilemmas—controlled cognitive processes occur often in response to dilemmas such as the trolley dilemma and automatic emotional responses occur often in response to dilemmas such as the footbridge dilemma [30]. Thus, in the trolley dilemma, controlled cognitive processes may drive people's attention to the agent type and make the judgment that it is inappropriate for AI agents to make moral decisions. In the footbridge dilemma, the action of pushing someone off a footbridge may evoke a stronger negative emotion than the action of operating a switch in the trolley dilemma. Driven by these automatic negative emotional responses, people would focus more on whether the agents did this harmful act, and judged this harmful act less acceptable and more morally wrong.

However, it should be noted that our work presents some limitations and offers several avenues for future research. Firstly, the current study only examined how people make moral judgments about humans and AI agents, but did not investigate the underlying psychological mechanism. Thus, all interpretations of the results are speculations. Future research could further explore the reason why people are reluctant to AI agents making moral decisions in the trolley dilemma, why people apply the same moral norms to humans and AI agents in the footbridge dilemma, and why people show different patterns of moral judgment in the trolley dilemma and the footbridge dilemma. Previous research has provided us with some pointers. For example, interpretability and consistency of behaviors would increase people's acceptance of AI [50,51]; increased anthropomorphism of an autonomous agent would mitigate blame for an agent's involvement in an undesirable outcome [52]. Individual

differences including personality [53,54], development experiences [55,56,57], and cultural background [58] may also influence people's attitudes toward AI agents. Second, to exclude the potential influence of individual differences between Experiments 1 and 2, we conducted Experiment 3 with a within-subjects design, participants were asked to read both the two scenarios; however, the processing system activated by the first scenario may influence the participants' judgment about the subsequent scenario. For example, participants who read the footbridge dilemma first may be interested in whether the character acted or not due to the strong negative emotion, this emotion may drive people to focus on the character's action in the subsequent trolley dilemma, just like they did in the footbridge dilemma. Future research could consider other method approaches to exclude the effects of individual differences and order effects.

Last but not least, we examined people's moral judgments of humans and AI agents using the most classical and traditional dilemma paradigms. However, recent studies have discussed the method limitations of these traditional dilemma paradigms. For example, Gawronski et al. [59] claimed that the conceptual meaning of responses in the traditional paradigm is ambiguous because the central aspects of utilitarianism and deontology—consequences and norms—are not manipulated. This shortcoming may undermine empirical findings such as the problem that the traditional moral paradigms could not measure the general action/inaction biases of participants. Given this, there might be other explanations for the current results, such as the inclination to be overall acceptable to any behavioral proposals. Specifically, there might be a stronger overall inclination to accept the action choices in the trolley dilemma than in the footbridge dilemma. This is an alternative explanation that should be tested in future studies. To resolve these limitations of traditional dilemma paradigms, Gawronski et al. presented a multinomial model (the CNI model) that allows researchers to quantify sensitivity to consequences, sensitivity to moral norms, and a general preference for inaction versus action irrespective of consequences and norms in responses to moral dilemmas (for details, see [59,60]). More recently, however, the limitations of the CNI model were also been discussed (for details, see [61,62]). As a result, Liu and Liao [63,64] developed a new algorithm—the CAN—to address the methodological limitations of the CNI model and fix these limitations (for details, see [63,64,65]). We suggest that further research could examine people's moral judgments

of humans and AI agents using these newly developed approaches and make a comparison with the current results that come from the traditional dilemma paradigms.

Author Contributions

Y.Z. implemented the experiments, collected, analyzed, and interpreted data, wrote the "Materials and Methods" and "Results" of the manuscript; J.W. analyzed and interpreted data, wrote the introduction and discussion of the manuscript; F.Y. described the proposed framework and revised the manuscript; L.X. supervised and guided the experiments, revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki. The studies involving human participants were reviewed and approved by the Ethics Committee of Wuhan University.

Informed Consent Statement

Informed consent was obtained from all participants involved in the study.

Data Availability Statement

The authors will share data from the study upon reasonable request to the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

Funding Statement

This research was supported by the National Social Science Foundation of China (Grant No. 20CZX059), and the National Natural Science Foundation of China (Grant No. 72101132).

Footnotes

References

1. Rabbitt S.M., Kazdin A.E., Scassellati B. Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. Clin. Psychol. Rev. 2015;35:35–46. doi: 10.1016/j.cpr.2014.07.001. [DOI] [PubMed] [Google Scholar]

2. Fournier T. Will my next car be a libertarian or a utilitarian? Who will decide? IEEE Veh. Technol. Mag. 2016;35:40–45. doi: 10.1109/MTS.2016.2554441. [DOI] [Google Scholar]

3. Angwin J., Larson J., Surya M., Lauren K. Machine Bias. [(accessed on 30 December 2022)]. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminalsentencing.

4. Moulliet D., Stolzenbach J., Majonek A., Völker T. The expansion of Robo-Advisory in Wealth Management. [(accessed on 30 December 2022)]. Available online: https://www2.deloitte.com/content/dam/Deloitte/de/Documents/financialservices/Deloitte-Robo-safe.pdf.

5. Shank D.B., DeSanti A., Maninger T. When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. Inf. Commun. Soc. 2019;22:648–663. doi: 10.1080/1369118X.2019.1568515. [DOI] [Google Scholar]

6. Malle B.F., Scheutz M., Arnold T., Voiklis J., Cusimano C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents; Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI); Portland, OR, USA. 2–

5 March 2015; pp. 117–124. [Google Scholar]

7. Voiklis J., Kim B., Cusimano C., Malle B.F. Moral judgments of human vs. robot agents; Proceedings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN); New York, NY, USA. 26–31 August 2016; pp. 775–780. [Google Scholar]

8. Bigman Y.E., Gray K. People are averse to machines making moral decisions. Cognition. 2018;181:21–34. doi: 10.1016/j.cognition.2018.08.003. [DOI] [PubMed] [Google Scholar]

9. Crowley J. Woman Says Amazon's Alexa Told Her to Stab Herself in The Heart For 'The Greater Good'. [(accessed on 30 December 2022)]. Available online: https://www.newsweek.com/amazon-echo-tells-uk-woman-stab-herself-1479074.

10. Forrest C. Robot Kills Worker on Assembly Line, Raising Concerns about Human-Robot Collaboration. [(accessed on 30 December 2022)]. Available online: https://www.techrepublic.com/article/robot-kills-worker-on-assembly-line-raising-concerns-about-human-robot-collaboration/

11. Schwab K. The Fourth Industrial Revolution. [(accessed on 30 December 2022)]. Available online: https://books.google.com/books?hl=en&lr=&id=GVekDQAAQBAJ&oi=fnd&pg=PR7&dq=The+fourth+industrial+revolution&ots=NhKeFDzwhG&sig=SxKMGj8OWFndH_0YSdJMKbknCwA#v=onepage&q=The+fourth+industrial+revolution&f=false.

12. Komatsu T., Malle B.F., Scheutz M. Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across US and Japan; Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction; Boulder, CO, USA. 8–11 March 2021; pp. 63–72. [Google Scholar]

13. Lee M.K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data Soc. 2018;5:1–16. doi: 10.1177/2053951718756684. [DOI] [Google Scholar]

14. Longoni C., Bonezzi A., Morewedge C.K. Resistance to medical artificial intelligence. J. Consum. Res. 2019;46:629–650. doi: 10.1093/jcr/ucz013. [DOI] [Google Scholar]

15. Malle B.F. Moral judgments. Annu. Rev. Psychol. 2021;72:293–318. doi: 10.1146/annurev-

psych-072220-104358. [DOI] [PubMed] [Google Scholar]

16. De Houwer J., Thomas S., Baeyens F. Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. Psychol. Bull. 2001;127:853–869. doi: 10.1037/0033-2909.127.6.853. [DOI] [PubMed] [Google Scholar]

17. Cusimano C., Thapa S., Malle B.F. Judgment before emotion: People access moral evaluations faster than affective states; Proceedings of the 39th Annual Conference of the Cognitive Science Society; London, UK. 26–29 July 2017; pp. 1848–1853. [Google Scholar]

18. Niedenthal P.M., Rohmann A., Dalle N. What is primed by emotion concepts and emotion words? In: Musch J., Klauer K.C., editors. The Psychology of Evaluation: Affective Processes in Cognition and Emotion. Erlbaum; Mahwah, NJ, USA: 2003. pp. 307–333. [Google Scholar]

19. Nichols S., Mallon R. Moral dilemmas and moral rules. Cognition. 2006;100:530–542. doi: 10.1016/j.cognition.2005.07.005. [DOI] [PubMed] [Google Scholar]

20. Cosmides L., Tooby J. Cognitive adaptations for social exchange. In: Barkow J., Cosmides L., Tooby J., editors. The Adapted Mind. Oxford University Press; New York, NY, USA: 1992. pp. 163–228. [Google Scholar]

21. Stone V.E., Cosmides L., Tooby J., Kroll N., Knight R.T. Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. Proc. Natl. Acad. Sci. USA. 2002;99:11531–11536. doi: 10.1073/pnas.122352699. [DOI] [PMC free article] [PubMed] [Google Scholar]

22. Gigerenzer G., Hug K. Domain-specific reasoning: Social contracts, cheating, and perspective change. Cognition. 1992;43:127–171. doi: 10.1016/0010-0277(92)90060-U. [DOI] [PubMed] [Google Scholar]

23. Haidt J. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychol. Rev. 2001;108:814–834. doi: 10.1037/0033-295X.108.4.814. [DOI] [PubMed] [Google Scholar]

24. Alicke M.D. Culpable control and the psychology of blame. Psychol. Rev. 2000;126:556–574. doi: 10.1037/0033-2909.126.4.556. [DOI] [PubMed] [Google Scholar]

25. Malle B.F., Guglielmo S., Monroe A.E. A theory of blame. Psychol. Inq. 2014;25:147–186. doi:

10.1080/1047840X.2014.877340. [DOI] [Google Scholar]

26. Mill J.S. Utilitarianism. In: Robson J.M., editor. Essays on Ethics, Religion, and Society. Volume 10. University of Toronto Press; Toronto, ON, Canada: 1969. pp. 203–259. [Google Scholar]

27. Kant I. Deontology: The Ethics of Duty. In: Keller D., editor. Ethics and Values: Basic Readings in Theory and Practice. Pearson Custom; Boston, MA, USA: 2002. pp. 77–87. [Google Scholar]

28. Greene J.D., Sommerville R.B., Nystrom L.E., Darley J.M., Cohen J.D. An fMRI investigation of emotional engagement in moral judgment. Science. 2001;293:2105–2108. doi: 10.1126/science.1062872. [DOI] [PubMed] [Google Scholar]

29. Greene J.D. Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. Trends Cogn. Sci. 2007;11:322–323. doi: 10.1016/j.tics.2007.06.004. [DOI] [PubMed] [Google Scholar]

30. Greene J.D. Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. J. Exp. Soc. Psychol. 2009;45:581–584. doi: 10.1016/j.jesp.2009.01.003. [DOI] [Google Scholar]

31. Greene J.D., Morelli S.A., Lowenberg K., Nystrom L.E., Cohen J.D. Cognitive load selectively interferes with utilitarian moral judgment. Cognition. 2008;107:1144–1154. doi: 10.1016/j.cognition.2007.11.004. [DOI] [PMC free article] [PubMed] [Google Scholar]

32. Cushman F., Young L., Hauser M. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. Psychol. Sci. 2006;17:1082–1089. doi: 10.1111/j.1467-9280.2006.01834.x. [DOI] [PubMed] [Google Scholar]

33. Schaich Borg J., Hynes C., Van Horn J., Grafton S., Sinnott-Armstrong W. Consequences, action, and intention as factors in moral judgments: An fMRI investigation. J. Cogn. Neurosci. 2006;18:803–817. doi: 10.1162/jocn.2006.18.5.803. [DOI] [PubMed] [Google Scholar]

34. Royzman E.B., Baron J. The preference for indirect harm. Soc. Justice Res. 2002;15:165–184. doi: 10.1023/A:1019923923537. [DOI] [Google Scholar]

35. Waldmann M.R., Dieterich J.H. Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. Psychol. Sci. 2007;18:247–253. doi: 10.1111/j.1467-9280.2007.01884.x. [DOI] [PubMed] [Google Scholar]

36. Greene J.D., Cushman F.A., Stewart L.E., Lowenberg K., Nystrom L.E., Cohen J.D. Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition. 2009;111:364–371. doi: 10.1016/j.cognition.2009.02.001. [DOI] [PubMed] [Google Scholar]

37. Yagoda R.E., Gillan D.J. You want me to trust a robot? The development of a human-robot interaction trust scale. Int. J. Soc. Robot. 2002;4:235–248. doi: 10.1007/s12369-012-0144-0. [DOI] [Google Scholar]

38. Naneva S., Sarda Gou M., Webb T.L., Prescott T.J. A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. Int. J. Soc. Robot. 2020;12:1179–1201. doi: 10.1007/s12369-020-00659-4. [DOI] [Google Scholar]

39. Fortunati L., Sarrica M., Ferrin G., Brondi S., Honsell F. Social robots as cultural objects: The sixth dimension of dynamicity? Inf. Soc. 2018;34:141–152. doi: 10.1080/01972243.2018.1444253. [DOI] [Google Scholar]

40. Gray H., Gray K., Wegner D.M. Dimensions of mind perception. Science. 2007;315:619. doi: 10.1126/science.1134475. [DOI] [PubMed] [Google Scholar]

41. Gray K., Wegner D.M. Feeling robots and human zombies: Mind perception and the uncanny valley. Cognition. 2012;125:125–130. doi: 10.1016/j.cognition.2012.06.007. [DOI] [PubMed] [Google Scholar]

42. Gray K., Young L., Waytz A. Mind perception is the essence of morality. Psychol. Inq. 2012;23:101–124. doi: 10.1080/1047840X.2012.651387. [DOI] [PMC free article] [PubMed] [Google Scholar]

43. Newcomb D. You Won't Need a Driver's License by 2040. [(accessed on 14 January 2023)]. Available online: http://edition.cnn.com/2012/09/18/tech/innovation/ieee-2040-cars.

44. Van Arem B., Van Driel C.J., Visser R. The impact of cooperative adaptive cruise control on traffic-flow characteristics. IEEE Trans. Intell. Transp. Syst. 2006;7:429–436. doi: 10.1109/TITS.2006.884615. [DOI] [Google Scholar]

45. Gao P., Hensley R., Zielke A. A roadmap to the future for the auto industry. [(accessed on 30 December 2022)]. Available online: https://img.etb2bimg.com/files/retail_files/reports/data_file-A-road-map-to-the-future-for-

the-auto-industry-McKinsey-Quarterly-Report-1426754280.pdf.

46. Bonnefon J.F., Shariff A., Rahwan I. The social dilemma of autonomous vehicles. Science. 2016;352:1573–1576. doi: 10.1126/science.aaf2654. [DOI] [PubMed] [Google Scholar]

47. Komatsu T. Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds; Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI); Christchurch, New Zealand. 7–10 March 2016; pp. 457–458. [Google Scholar]

48. Thomson J.J. Killing, letting die, and the trolley problem. Monist. 1976;59:204–217. doi: 10.5840/monist197659224. [DOI] [PubMed] [Google Scholar]

49. Foot P. The problem of abortion and the doctrine of the double effect. Oxford. Rev. 1967;2:152–161. [Google Scholar]

50. Salem M., Lakatos G., Amirabdollahian F., Dautenhahn K. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust; Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction; Portland, OR, USA. 2–5 March 2015; pp. 141–148. [Google Scholar]

51. Robinette P., Howard A.M., Wagner A.R. Effect of Robot Performance on Human Robot Trust in Time-Critical Situations. IEEE Trans. Hum. Mach. Syst. 2017;47:425–436. doi: 10.1109/THMS.2017.2648849. [DOI] [Google Scholar]

52. Waytz A., Heafner J., Epley N. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. 2014;52:113–117. doi: 10.1016/j.jesp.2014.01.005. [DOI] [Google Scholar]

53. Ho G., Wheatley D., Scialfa C.T. Age differences in trust and reliance of a medication management system. Interact. Comput. 2015;17:690–710. doi: 10.1016/j.intcom.2005.09.007. [DOI] [Google Scholar]

54. May D.C., Holler K.J., Bethel C.L., Strawderman L., Carruth D.W., Usher J.M. Survey of factors for the prediction of human comfort with a non-anthropomorphic robot in public spaces. Int. J. Soc. Robot. 2017;9:165–180. doi: 10.1007/s12369-016-0390-7. [DOI] [Google Scholar]

55. Haring K.S., Matsumoto Y., Watanabe K. How do people perceive and trust a lifelike robot;

Proceedings of the World Congress on Engineering and Computer Science; San Francisco, CA, USA. 23–25 October 2013; pp. 425–430. [Google Scholar]

56. Joosse M., Lohse M., Perez J.G., Evers V. What you do is who you are: The role of task context in perceived social robot personality; Proceedings of the 2013 IEEE International Conference on Robotics and Automation; Karlsruhe, Germany. 6–10 May 2013; pp. 2134–2139. [Google Scholar]

57. Niculescu A., van Dijk B., Nijholt A., Li H., See S.L. Making social robots more attractive: The effects of voice pitch, humor and empathy. Int. J. Soc. Robot. 2013;5:171–191. doi: 10.1007/s12369-012-0171-x. [DOI] [Google Scholar]

58. Huerta E., Glandon T., Petrides Y. Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. Int. J. Appl. Inf. Syst. 2012;13:316–333. doi: 10.1016/j.accinf.2012.03.007. [DOI] [Google Scholar]

59. Gawronski B., Armstrong J., Conway P., Friesdorf R., Hütter M. Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. J. Pers. Soc. Psychol. 2017;113:343–376. doi: 10.1037/pspa0000086. [DOI] [PubMed] [Google Scholar]

60. Gawronski B., Beer J.S. What makes moral dilemma judgments “utilitarian” or “deontological”? Soc. Neurosci. 2017;12:626–632. doi: 10.1080/17470919.2016.1248787. [DOI] [PubMed] [Google Scholar]

61. Baron J., Goodwin G.P. Consequences, norms, and inaction: A critical analysis. Judgm. Decis. 2020;15:421–442. doi: 10.1017/S193029750000721X. [DOI] [Google Scholar]

62. Baron J., Goodwin G.P. Consequences, norms, and inaction: Response to Gawronski et al. (2020) Judgm. Decis. 2021;16:566–595. doi: 10.1017/S1930297500008676. [DOI] [Google Scholar]

63. Liu C., Liao J. CAN algorithm: An individual level approach to identify consequence and norm sensitivities and overall action/inaction preferences in moral decision-making. Front. Psychol. 2021;11:547916. doi: 10.3389/fpsyg.2020.547916. [DOI] [PMC free article] [PubMed] [Google Scholar]

64. Liu C., Liao J. Stand up to action: The postural effect of moral dilemma decision-making and the moderating role of dual processes. PsyCh J. 2021;10:587–597. doi: 10.1002/pchj.449. [DOI]

[PubMed] [Google Scholar]

65. Feng C., Liu C. Resolving the Limitations of the CNI Model in Moral Decision Making Using the CAN Algorithm: A Methodological Contrast. Behav. Sci. 2022;12:233. doi: 10.3390/bs12070233. [DOI] [PMC free article] [PubMed] [Google Scholar]

Associated Data

This section collects any data citations, data availability statements, or supplementary materials included in this article.

Data Availability Statement

The authors will share data from the study upon reasonable request to the corresponding author.