Journal of Experimental Social Psychology 101 (2022) 104327

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI☆

Zaixuan Zhang [a], Zhansheng Chen [a,*], Liying Xu [b]

[a] Department of Psychology, The University of Hong Kong, Hong Kong
[b] Department of Psychology, School of Social Sciences, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
Artificial intelligence
Moral dilemmas
Warmth perception

ABSTRACT

Artificial intelligence (AI) has become deeply integrated into daily life; thus, it is important to examine how people perceive AI as it functions as a decision-maker, especially in situations involving moral dilemmas. Across four studies ($N = 804$), we found that people perceive AI as more likely to make utilitarian choices than human beings are (Studies1–4). We then measured people's perceptions (both warmth and competence) toward AI and explored their potential contributions to our predicted main effect (Study 2). In addition, our main effect was replicated in impersonal moral dilemma and personal high-conflict moral dilemma situations (Studies 3 and 4). We discuss the implications of these findings on moral dilemma and human–computer interactions.

Artificial intelligence (AI) is becoming increasingly integrated into daily life. It is critically important to examine how people perceive and judge AI's various acts. For instance, imagine an AI-controlled self-driving car is rushing toward a passerby, and there is something wrong with its brake pad. The car has the option of switching to another path, but this will result in hitting another five people on the other path. How would an AI act in a situation such as this? In this research, we focused on how people would perceive an AI's behavior or behavioral tendencies in situations involving moral dilemmas. The potential mechanism and moderators of the relationship were also explored.

1. Perception of artificial intelligence

AI is gaining increasing popularity in the domains of human life, including economic decision-making (Gogoll & Uhl, 2018), the game of GO (Silver, Huang, Maddison, Guez, & Hassabis, 2016), criminal sentencing (Hayashi & Wakabayashi, 2018), military actions on the battlefield (Horowitz, 2016), etc. Yet, how people perceive and predict AI's behaviors is not well studied. In retrospect, the computer was the first artificial agent that drew social scientists' attention. According to past research, individuals tended to personify computers even when the computers were not like humans (Nass, Moon, Fogg, Reeves, & Dryer, 1995). The same group of scholars also found that human and computer teammates were perceived as similar in a "Desert Survival Task" (Nass, Fogg, & Moon, 1996).

Human beings' mind perceptions can offer insights for understanding how people perceive AI. Mind perception involves two dimensions: agency and experience (Gray, Gray, & Wegner, 2007). Agency refers to the capacity to execute a task, whereas experience relates to the ability to feel, especially concerning emotions. Gray and Wegner (2012) argued that these two dimensions related strongly, both theoretically and empirically, to the content of stereotype (warmth/competence) that Fiske, Cuddy, and Glick (2007) proposed. Besides, there is often a negative dynamic relationship between the perceived warmth and competence of a given entity (Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). Empirical evidence showed that robots were perceived with low warmth but high competence (Liu, Shen, & Hancock, 2021), even those robots designed for social interaction with human beings. Although these scholars did not provide further explanation for this finding, it may be due to AI's high efficiency (e.g., fast calculation) but undetectable intentions (e.g., over-complex algorithm). It is reasonable to predict that an AI might be considered as lower in warmth and higher in competence when compared to humans.

The literature has accumulated evidence that reflects people's perception of AI (i.e., high in competence/agency and low in warmth/experience). For instance, Gogoll and Uhl (2018) found that individuals tended to be hesitant when delegating a task to an AI, because they generally holding a relatively critical opinion of such delegation.

☆ This paper has been recommended for acceptance by Lasana Harris.

* Corresponding author at: Department of Psychology, The University of Hong Kong.
E-mail address: chenz@hku.hk (Z. Chen).

https://doi.org/10.1016/j.jesp.2022.104327
Received 26 June 2021; Received in revised form 13 January 2022; Accepted 11 March 2022
Available online 5 April 2022

However, when an AI served as only an adviser in general issues (e.g., selecting songs), people tended to adhere to its decision (Logg, Minson, & Moore, 2019).

In addition, Castelo, Bos, and Lehmann (2019) found that people trust algorithms (which are inside the computer and AI) more when they dealt with objective rather than subjective tasks. Regarding ethical issues, people tend to rule out AI for morality-related decisions (e.g., medical decisions; Longoni, Bonezzi, & Morewedge, 2019), while this tendency decreases with increased individual exposure to AI (Kramer, Schaich Borg, Conitzer, & Sinnott-Armstrong, 2018). Further, when there is some moral wrongness, people are more likely to attribute it to AI (Shank & Desanti, 2018) because they judge AI as more blameworthy and less moral compared to humans (Young & Monroe, 2019). Nevertheless, given the increasingly significant role that AI can play in daily life, more studies that are empirical are needed to address people's predicted actions of AI in situations involving ethical decision-making.

2. AI in moral dilemma

Morality, especially moral judgment and decision-making, can be debatable and ambiguous (Dunning, Meyerowitz, & Holzberg, 1989). In dealing with moral dilemmas, a given decision can be classified as utilitarian versus deontological (Carmona-Perera, Caracuel, Pérez-García, & Verdejo-García, 2015). While the utilitarian approach insists on accepting harm and focuses on outcomes (Mill, 1861/1998), the deontological approach rejects any harm and concentrates on the nature of the moral action (Kant, 1785/1959). The well-known "Trolley Dilemma" (Foot, 1967) was created to contrast these two approaches. Subsequently, other abstract dilemmas emerged, such as the "Foot Bridge Dilemma" (Thomson, 1986). Returning to the above scenario regarding AI-controlled self-driving cars, people's predictions about the AI's behavior would affect not only how they would apply AI in making such decisions, but also their attitudes toward the so-called "fact" and its consequences.

People can infer others' decisions and behaviors in moral dilemmas according to perceived warmth and competence (Rom & Conway, 2018). In particular, Rom, Weiss, and Conway (2017) found that people predicted that affective decision-makers (high warmth but low competence when facing moral dilemmas) would reject harm and make deontological decisions, whereas cognitive decision-makers (low warmth but high competence) would accept harm and make utilitarian decisions. They theorized that decision-makers with high warmth would be less empathic and experience more affective reactions toward the potential active harm, while decision-makers with high competence were perceived to be more outcome focused. Based on that, individuals may make a particular prediction about their decision in moral dilemmas as aforementioned.

According to the above, as people tend to unconsciously apply social rules and expectations from human society to computers and other IT devices (Nass & Moon, 2000), the way people perceive AI's warmth and competence may also influence the way they predict AI's behaviors in moral dilemmas. Specifically, because of its skilled power and mysterious inner workings, AI is more likely to be perceived as low warmth and high competence when compared to humans. Regarding this, AI would then be perceived as an entity with relatively less empathy and have a more outcome orientation. Thus, we hypothesized that, relative to human beings, people would predict AI to behave in a more utilitarian manner (high harm acceptance).

Moreover, moral dilemmas can be categorized into three types: personal high-conflict, personal low-conflict, and impersonal (Greene, Nystrom, Engell, Darley, & Cohen, 2004). Generally, in personal moral dilemmas, the agent's force directly causes the harm. In impersonal moral dilemmas, forces other than the agents (e.g., guns, trolleys, or disease; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008) cause the harms. Personal moral dilemmas are dealt with more social-emotional responses because decision-makers would induce the harm directly on their own. In contrast, impersonal moral dilemmas require more cognitive processes, partly because the decision-makers do not cause the harm with their own hands. High and low conflict indicated different degrees of emotional engagement (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Personal high-conflict moral dilemmas always induce the strongest aversion, with nearly 90% of respondents unwilling to make a utilitarian decision in that occasion (Cushman, Young, & Hauser, 2006). The type of moral dilemma may influence how people predict AI's decisions. Specifically, as the personal high-conflict moral dilemmas elicit the strongest emotional reaction, human beings perceived to be more affective (vs. AI) would then be predicted to make the fewest utilitarian decisions in that condition, while people's prediction toward AI may not be affected due to its lower perceived empathy. Thus, we hypothesize that although an AI will be perceived as more utilitarian in moral dilemmas than humans would be, the magnitude of its perceived utilitarian decision may vary regarding the type of moral dilemmas.

3. The current research

In the current research, we aimed to examine how people perceive AI's action in moral dilemmas and to investigate the generalization of this perception in different contexts. First, we tested whether people perceived AI as being more likely to make utilitarian choices in moral dilemmas than human beings are (Study 1) and whether a lower level of perceived warmth in AI could be a potential internal mechanism (Study 2). We then tested our predicted effect across three types of moral dilemmas (personal high-conflict, personal low-conflict, and impersonal), and we expected a replication of the main effect (Study 3). Finally, in Study 4, we replicated Study 3 with participants from the United States.

4. Study 1

In this study, we explored people's inferences concerning the identity of the decision-makers in two different conditions: participants were confronted with a decision-maker who made either a utilitarian or a deontological choice in the trolley dilemma. Then, participants inferred the identity of the decision-maker (AI vs. human). Because participant were inferring the probablity of two possible outcomes they were asked to indicate the possibility of each decision-maker (the sum of possibilities was always 100%), which made the dependent variable continuous. Besides, because there were only two potential choices, the more possibility signed to "AI" meant less possibility signed to "human." In other words, these two items mirrored each other. Regarding this, we only involved the possibility signed to "AI" in the data analysis. In addition, as individuals' existing attitudes can influence or bias their perception and judgments (Ajzen & Fishbein, 2005), there could also be an alternative that participants' own tendency in the original trolley dilemma might affect their prediction on the decision-makers' identity. Hence, we also tested participants' own tendency in the original trolley dilemma as an individual difference and explore its potential effect.

4.1. Participants and design

In total, 150 Chinese college students participated in exchange for 2 RMB (around 0.3 USD). We excluded 32 for failing the attention check. A sensitivity test (Faul, Erdfelder, Lang, & Buchner, 2007) showed that our final sample size of 122 (73 females, $M_{age} = 22.09$, $SD_{age} = 2.58$) could provide 80% power to detect an effect of $d = 0.45$ (medium; $\alpha = 0.05$). We randomly assigned participants to one of two experimental conditions (utilitarian choice condition vs. deontological choice condition).

4.2. Procedure and measures

All participants were first presented with a standard trolley dilemma

problem, and they were asked to report their own behavioral tendencies in an original trolley dilemma (i.e., "Would you switch the track to save five people with a single casualty, or stand by and do nothing?"). Next, participants were shown one of two figures (see Fig. 1). In both figures, the participant was a passerby who could do nothing, and there was a control room beside the railroad switch. Importantly, in one figure, the track was switched and the trolley was going to the single person (i.e., the utilitarian decision condition); in the other, the track was not switched and the trolley was approaching the five people (i.e., the deontological decision condition). Then, participants were told there was an entity in the control room who had just made the decision and that it could be an AI system or a human. The participants were asked to indicate the percent possibility of each potential decision-maker.

4.3. Result and discussion

An independent-sample *t*-test revealed a significant discrepancy in perceived decision-maker identity between the utilitarian condition ($M = 62.93$, $SD = 28.43$) and the deontological condition ($M = 47.29$, $SD = 35.11$), $t = -2.71$, $p < .01$, $d = 0.49$, 95% CI = [−27.07, −4.23]. That is, when a utilitarian choice was made, relative to when a deontological choice was made, participants believed more strongly that the decision-maker was an AI system.

At the beginning of the standard trolley problem, 58 participants (47.5%) indicated that they would make utilitarian choices, while the other 64 participants (52.5%) tended to make deontological choices. To examine whether this individual difference among participants could influence their inference of the second scenario, we conducted a two-way analysis of variance (ANOVA) on the experimental condition participants faced and their moral judgment in the original trolley dilemma. The main effect of the experimental condition was significant, $F(1,118) = 7.54$, $p < .01$, $\eta_p^2 = 0.06$. However, the two-way interaction was not significant, $F(1,118) = 0.34$, $p = .56$, $\eta_p^2 = 0.003$, let alone the main effect of the individual difference, $F(1,118) = 2.60$, $p = .11$, $\eta_p^2 = 0.022$. Thus, participants' original opinions in the standard trolley problem did not influence their responses in the second scenario.

Study 1 provided initial support to our hypothesis. Specifically, individuals tended to believe that the utilitarian decision-maker was more likely to be an AI than a human being was, which was consistent with our main hypothesis. Peoples' opinions in the original moral dilemma did not influence this effect. Nevertheless, it remained unclear why people held this attitude regarding AI versus human beings, which was addressed in the next study.

5. Study 2

Instead of judging the likelihood of an AI versus a human being committing an observed action, as was the case in Study 1, participants in Study 2 were asked to judge a given protagonist's likelihood of engaging in a utilitarian choice under an imagined scenario. The protagonist was either an AI or a human being. In addition, how the participants perceived the protagonist's warmth and competence was assessed to test these factors' potential contributions to peoples' predicted differences in moral decisions between AI and human beings.

5.1. Participants and design

Same as in Study 1, 150 college students from China participated in exchange for 2 RMB. We excluded 18 participants for failing the attention check. A sensitivity test (Faul et al., 2007) showed that our final sample size of 132 (74 females, $M_{age} = 21.04$, $SD_{age} = 2.07$) could provide 80% power to detect an effect of $d = 0.43$ (small to medium; $\alpha = 0.05$). We randomly assigned participants to one of two decision-maker conditions (AI system vs. human being).

5.2. Procedure and measures

Participants in the two experimental conditions were shown different figures similar to those in Study 1 (see supplementary materials for details). In both figures, the participant was a passerby who could do nothing, and there was a control room beside the railroad switch with a "decision-maker" inside. Participants were told the identity of the decision-maker in the control room (AI system vs. human). Then, the participants were asked to report their perceptions, including warmth (i.e., kind, friendly, sincere) and competence (i.e., smart, efficient, wise), toward the decision-maker on a 7-point scale (e.g., "To what extent do you think the 'decision-maker is kind?"; 1 = *not at all*, 7 = *extremely*). Finally, participants were asked to indicate the percent possibility that the decision-maker would make a utilitarian choice.

5.3. Result and discussion

An independent-sample *t*-test revealed a significant main effect on perceived likelihood of making a utilitarian choice, $t = -2.16$, $p < .05$, $d = 0.38$, 95% CI = [−19.6, −1.02]; such that participants perceived the AI system ($M = 77.32$, $SD = 25.23$) as having a higher chance of making a utilitarian choice than the human being would ($M = 67.01$, $SD = 28.56$). This indicated that if the decision-maker was an AI system, participants preferred to believe that the decision-maker would make a utilitarian choice. Meanwhile, participants tended to believe that humans would be more likely to make a deontological choice. There was also a significant discrepancy between participants in two experimental conditions on their warmth perception toward a particular decision-maker, $t = 2.91$, $p < .01$, $d = 0.51$, 95% CI = [0.153, 0.857]. However, no effect on their competence perception, $t = -0.7$, $p > .05$, $d = -0.01$, 95% CI = [−0.353, 0.330]. That is, AI ($M = 4.22$, $SD = 0.94$) was

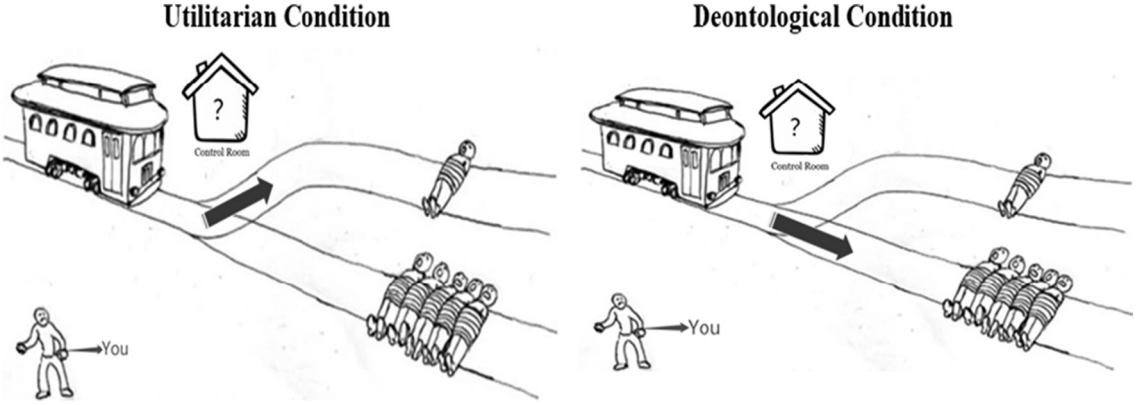

Fig. 1. The Two Experimental Conditions in Study 1.

perceived as less warm compared to human beings ($M = 4.66$, $SD = 0.80$) when serving as decision-makers, while they were perceived as equally competent.

To test our predicted role of perceived warmth for the above effect, we conducted a mediational analysis using PROCESS Model 4 (Hayes, 2012). The indirect effect of warmth perception was statistically significant, indirect effect = −3.82, $SE = 1.74$, 95%CI = [−7.23, −0.41], (Human = 0, AI = 1; 5000 iterations, bias corrected), which supported the mediating role of warmth. That is, when an AI functioned as the decision-maker in a moral dilemma scenario, participants perceived it as less warm, relative to a human being; thus, participants were more likely to predict that the AI would make a utilitarian choice. To test further this mechanism, we conducted another mediational analysis involving warmth perception as the outcome variable and the chance of utilitarian choice as the mediator. The analysis did not support the alternative mediational process, indirect effect = 0.07, $SE = 0.042$, 95%CI = [−0.013, 0.153], (Human = 0, AI = 1; 5000 iterations, bias corrected).

In addition, we tested the mediating role of the decision-maker's perceived competence using the PROCESS Model 4 (Hayes, 2012). The indirect effect of competence perception was not significant, indirect effect = −0.05, SE = 0.68, 95%CI = [−1.37, 1.28], (Human = 0, AI = 1; 5000 iterations, bias corrected). Thus, people's perceptions of a particular decision-maker's competence could not explain our reported main effect of the perceived difference between AI and human beings (see Fig. 2).

Consistent with Study 1, Study 2 converged to support our hypothesis that AI is perceived as more likely than humans are to make utilitarian choices in moral dilemmas. Study 2 also indicated that people's perceptions of particular decision-makers' warmth possibly mediated this effect. However, as we did not identify other potential mediators theoretically, the current mediation model could be only one of several models possible. In the above two studies, all experimental scenarios were adapted from the trolley dilemma. According to Greene et al. (2004, 2008), there are three types of moral dilemmas: personal high-conflict moral dilemma, personal low-conflict moral dilemma, and impersonal moral dilemma. In the next two studies, we aimed to test whether our observed effect could also be observed in all types of moral dilemmas.

6. Study 3

There are three different types of moral dilemmas, and AI can engage in them all. In this study, we explored the potential moderating effect of the types of moral dilemmas. Generally, if the effect is stable enough, then people will judge AI as more likely to make a utilitarian choice in all three dilemma types. To manipulate adequately the type of moral dilemma, we created scenarios involving organ transplantation modified after the transplant dilemma (Thomson, 1986).

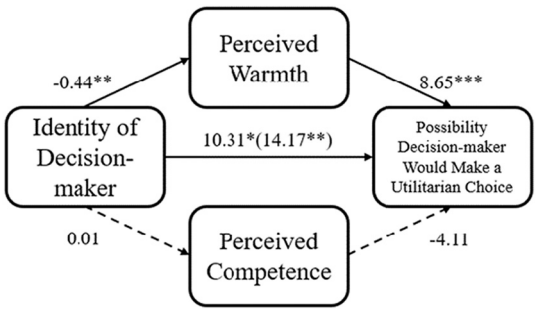

Fig. 2. Mediation Model of Decision-Makers' Identity on the Possibility that They Make a Utilitarian Choice, as Predicted via the Decision-Makers' Perceived Warmth.
Note. Human = 0, AI = 1; *p < .05. **p < .01. ***p < .001

6.1. Participants and design

In total, 247 Chinese college students participated in exchange for 2 RMB and certain credits. Of these, we excluded 28 for failing the attention check. A sensitivity test (Faul et al., 2007) revealed that our final sample size of 219 (127 females, $M_{age} = 21.59$, $SD_{age} = 2.42$) could provide 80% power to detect an effect of $\eta_p^2 = 0.042$ (small to medium; $\alpha = 0.05$) for a two-way ANOVA. We randomly assigned participants to one of six experimental conditions, and there was a 3 (type of dilemma: impersonal vs. personal high conflict vs. personal low conflict) × 2 (decision-maker: AI vs. human) between-subjects design.

6.2. Procedure and measures

The participants were first asked to read one of three moral dilemma scenarios. In the impersonal condition, participants read a scenario depicting one patient with multiple organ failure at the midpoint of their transplant surgery (two organs transplanted, two organs to be transplanted, and one under transplantation), and another five patients with single organ failure. There was a possible choice of stopping the current transplantation and saving the other five patients with these organs, but this would result in the first patient's death. In the personal high-conflict condition, participants read another scenario involving five patients with single organ failure and a healthy patient with all of their organs. There was a possible choice of saving the five patients at the cost of the healthy patient. In the personal low-conflict condition, participants read a scenario depicting five patients who need a particular medicine whose supply had run out for their treatment. There could be a choice to get the medicine from another patient but make the later patient suffer (see supplemental material for full text).

The participants were further informed that the decision-maker was either a human being or an AI. Participants were then asked to indicate the percent possibility that the decision-maker would make a utilitarian choice in their experimental conditions. Finally, the participants were thanked and debriefed.

6.3. Result and discussion

A two-way ANOVA revealed a significant main effect of decision-maker, $F(1,213) = 12.37$, $p < .001$, $\eta_p^2 = 0.055$; and a significant main effect of types of dilemma, $F(2,213) = 11.60$, $p < .001$, $\eta_p^2 = 0.098$; along with a significant interaction between the above two factors, $F(2,213) = 4.35$, $p < .05$, $\eta_p^2 = 0.039$. Given this, a least significant difference (LSD) test was conducted as well, which showed there were significant discrepancies between two decision-makers for impersonal, $t = 1.99$, $p < .05$, $d = 0.46$; and personal high-conflict dilemma, $t = 4.35$, $p < .001$, $d = 0.98$; but no significant discrepancy for personal low-conflict dilemma, $t = -0.02$, $p = .98$, $d = -0.01$ (Fig. 3). This indicated that people believed AI was more likely to make utilitarian choice in both impersonal and personal high-conflict moral dilemmas, but the two types of decision-makers were considered to make similar choices when confronted with the personal low-conflict dilemma (see Table 1).

According to our results, although participants in both impersonal and personal high-conflict dilemma conditions believed that AI was more likely to make a utilitarian choice (in line with our previous studies), there was no such effect for the personal low-conflict dilemma. That might result from the fact that a personal low-conflict dilemma could not stimulate enough emotional reaction when individuals were confronted with it. It was also possible that the effect in personal low-conflict dilemma was too small to be detected with the current sample size. That is, with the current sample size ($n = 32$ qualified participants in both conditions), only a large or effect or greater ($d > 0.62$) could be detected, according to a power analysis (Faul et al., 2007).

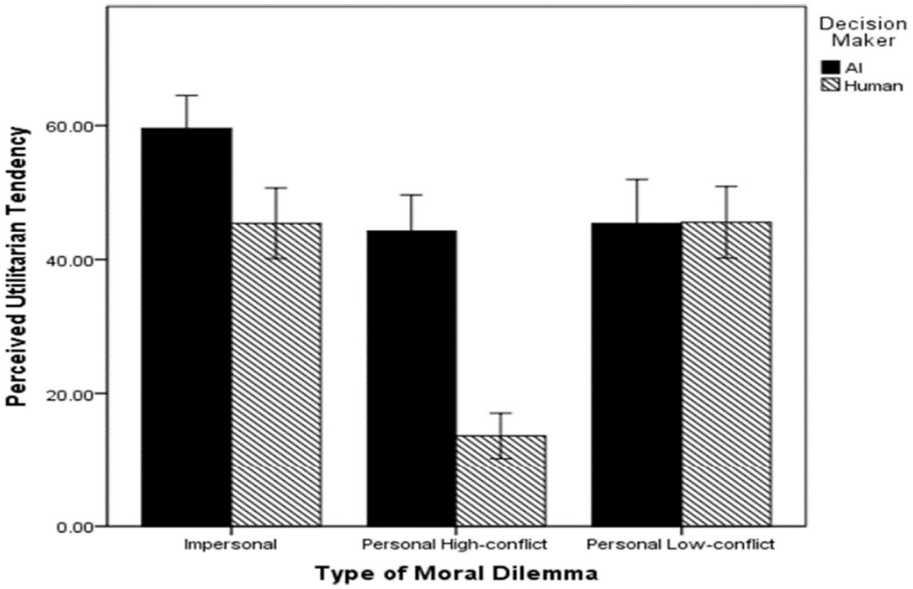

Fig. 3. Utilitarian Tendency toward Decision-Makers, as the Participants Perceived for Each Dilemma in Study 3.
Note. Error bars show standard errors

Table 1
Means (Standard Deviations) [Sample Size] of Participans' Perceived Utilitarian Tendency toward Decision-Makers in each Dilemma of Studies 3 and 4.

	Impersonal		Personal High-conflict		Personal Low-conflict	
	AI	Human	AI	Human	AI	Human
Perceived utilitarian tendency of DMs in Study 3	59.6 (31.9) [42]	45.4 (31.1) [35]	44.3 (33.6) [39]	13.6 (21.2) [39]	45.4 (37.2) [32]	45.6 (30.3) [32]
Perceived utilitarian tendency of DMs in Study 4	64.8 (26.3) [53]	55.0 (27.3) [52]	40.9 (29.3) [62]	26.1 (28.8) [57]	59.1 (27.7) [50]	60.1 (24.7) [57]

Note. DMs = Decision-Makers (i.e., AI or HUMAN).

7. Study 4

In Study 4, we aimed to replicate the effect in Study 3 with another sample for further reliability. The materials in this study were the same as in Study 3, except that the participants were recruited from the United States.

7.1. Participants and design

Using Amazon's Mechanical Turk (Rand, 2012), we recruited 370 American participants who were born and raised in the United States in exchange for 0.25 USD. We excluded 39 participants for failing the attention check. A sensitivity test (Faul et al., 2007) showed that our final sample size of 331 (173 females, $M_{age} = 40.49$, $SD_{age} = 12.55$) could provide 80% power to detect an effect of $\eta_p^2 = 0.023$ (small; $\alpha = 0.05$) for a two-way ANOVA. We randomly assigned participants to one of six experimental conditions, and there was a 3 (type of dilemma: impersonal vs. personal high conflict vs. personal low conflict) × 2 (decision-maker: AI vs. human) between-subjects design.

7.2. Procedure and measures

After being assigned to different experimental conditions, participants were asked to read the different moral dilemma scenarios. The material used in this study was the same as in Study 3. After that, participants were asked to indicate the percent possibility that the decision-maker would make a utilitarian choice in particular conditions.

7.3. Result and discussion

A two-way ANOVA revealed a significant main effect of the decision-maker, $F(1,325) = 6.78$, $p = .01$, $\eta_p^2 = 0.020$, and a significant main effect of the dilemma types, $F(2,325) = 34.84$, $p < .001$, $\eta_p^2 = 0.177$, along with a marginal significant interaction between the above two factors, $F(2,325) = 2.41$, $p = .092$, $\eta_p^2 = 0.015$. Given this, an LSD test was conducted as well, which showed that there were significant discrepancies between the two types of decision-makers for the personal high-conflict dilemma, $t = 2.94$, $p < .01$, $d = 0.54$; marginally significant discrepancy for the impersonal dilemma, $t = 1.84$, $p = .067$, $d = 0.36$; but no significant discrepancy for personal low-conflict dilemma, $t = -0.19$, $p = .85$, $d = -0.04$ (see Fig. 4). These results indicated that people held a very strong opinion that AI was more likely to make utilitarian choice in personal high-conflict moral dilemma, and a relatively weak opinion that AI was more likely to make utilitarian choice in the impersonal moral dilemma. However, the two decision-makers were expected to make similar choices when confronted with personal low-conflict dilemma.

In Studies 3 and 4, participants from diverse culture backgrounds had similar perceptions about AI's behavior in moral dilemmas. That is, both Chinese participants and American participants perceived that AI was more likely to make a utilitarian choice in moral dilemmas (i.e., impersonal and personal high-conflict moral dilemma). Besides, there were the largest discrepancies between AI and human beings for personal high-conflict condition for both Study 3 and Study 4 ($d = 0.98$ and 0.54, respectively), which was in line with what our prediction (see Table 2). Although different from what we observed in Study 3, there was only a marginally significant difference between AI and human beings in impersonal moral dilemma scenarios, and the effect size was reasonable ($d = 0.36$). In addition, the potential alternative discussed in Study 3 was not supported: even with an increased sample size, there was still no discrepancy observed in person low-conflict conditions ($p = .85$).

8. General discussion

AIs are becoming increasingly involved in daily life; however, there is limited knowledge on how people consider AI's moral behavior,

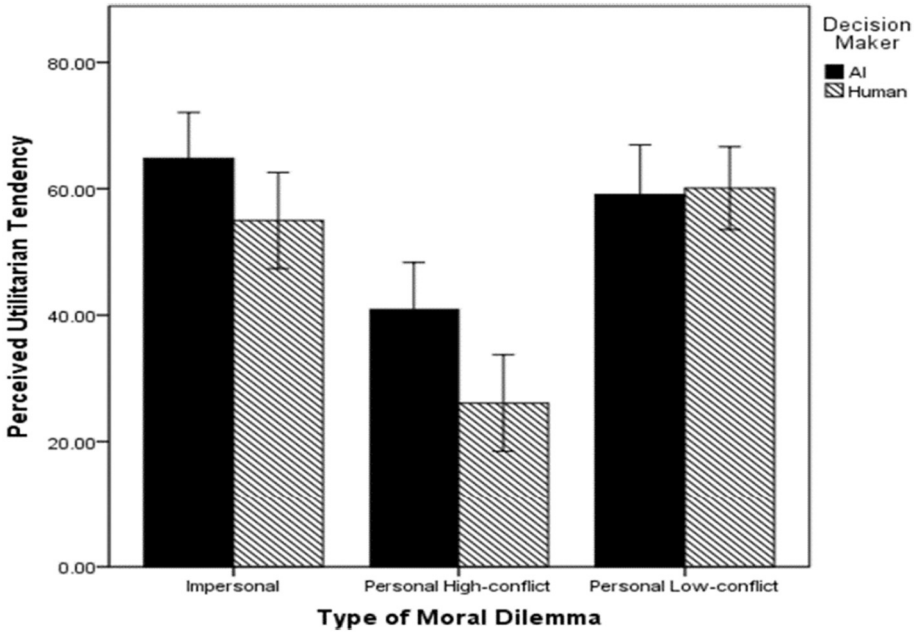

Fig. 4. Utilitarian Tendency of Decision-Makers, as the Participants Perceived for Each Dilemma in Study 4.
Note. Error bars show standard errors

Table 2
Details of the 2 × 3 Interaction in Studies 3 and 4.

	N	*F* values of Interaction	ES (η_p^2) of Interaction	ES (*d*) & 95%CI of Impersonal Condition	ES (*d*) & 95%CI of Hight-conflict Condition	ES (*d*) & 95%CI of Low-conflict Condition
Study 3	219	4.35	0.039	0.46 [0.002, 0.909]	0.98 [0.528, 1.44]	−0.01 [−0.499, 0.487]
Study 4	331	2.41	0.015	0.36 [−0.027, 0.743]	0.54 [0.176, 0.903]	−0.04 [−0.418, 0.344]

Note. ES = Effect Size.

especially in moral dilemmas involving moral conflicts. Across the four studies, we showed that AI was perceived to be more likely to make a utilitarian choice in moral dilemmas than humans would. Specifically, in Study 1, when a decision-maker made a utilitarian choice in the trolley dilemma, people tended to perceive the decision-maker as significantly more likely to be an AI than to be a human being. For Study 2, the effect was tested more directly. When the decision-maker for the moral dilemma was an AI, people believed it was more likely to make a utilitarian choice in the trolley dilemma, which may partially result from people's perception that AI would be relatively less warm. This effect could mostly be replicated across contexts, in that AI was thought to make utilitarian decisions more often, in both impersonal moral dilemmas and personal high-conflict moral dilemmas. However, there was no such effect for the personal low-conflict moral dilemma (Study 3). In Study 4, this effect was replicated with a larger sample from a different cultural background with the same method and procedure. Implications and limitations of the current research are discussed below.

8.1. Implications of the present research

Our investigation carries several implications. First, our reported differences between AI and human beings regarding people's predicted ethical decision-making enrich the literature on moral dilemma. AI is perceived to be more likely to make utilitarian choices than humans are in moral dilemmas. Besides, this effect may be induced by people's lower warmth perception toward AI. In line with previous research (Rom et al., 2017, Study 4), we also found that individuals who believed a less warm decision-maker was more likely be utilitarian in moral dilemma. That may result from the notion that a decision-maker with low warmth is less affective and harder to allow emotion to influence them, and then more likely to accept harm in the utilitarian choice. Although the same group of scholars (Rom et al., 2017, Study 1 and 2) suggested people also tend to perceive utilitarian decision-makers in moral dilemmas to be less warm, there is still no remarkable discrepancy between their finding and ours. This may due to two major reasons. First, individuals' social perception relies on what was firstly given. In the current research, it would be more direct to form perception according to the identities (AI vs. human) of the decision-makers, rather than basing the perception on the prediction toward their possible behavior (which their given identities also determine). Alternatively, the decision-makers' identities may cast a much stronger impact on individuals' perception than the predicted action of decision-makers does ($\eta_p^2 = 0.092$ vs. $\eta_p^2 = 0.049$) because in moral-related judgments or decision-making, who the actors are and what inside of them contribute more than what they actually do (Hechler & Kessler, 2018). Then, in the mediation model, the potential mediation effect of the predicted action is not significant. Thus, as we detected, perceived warmth should determine participants' prediction, rather than the participants' prediction determining it. In addition, researchers have already suggested several moderators to the decision-making process in moral dilemmas. For instance, individuals' decision-making systems (immediate action vs. long-range goals; Cushman, 2013) or ethical mindsets (rule based vs. outcome based; Cornelissen, Bashshur, Rode, & Le Menestrel, 2013) can influence this decision-making process. Given that previous research on moral dilemmas frequently assigned human beings as the decision-makers (e.g., Cushman, 2013; Rom & Conway,

2018), the main effect we detected suggested that the former effect could also differ when the decision-maker's identity was changed. For example, judging from the mediation process we discussed above, as homeless people are also perceived as being less warm (Fiske et al., 2007), individuals may blame them less even if they make utilitarian choices.

Second, our research supports the similarity between AI and human beings regarding the mechanism for people's predicted ethical decision-making. Study 2 showed individuals predicted that AI would be more likely to make utilitarian choices because of its lower warmth. This is consistent with Rom et al.' (2017) findings that a decision-maker with lower warmth will be perceived as more likely to be utilitarian.

Third, the current research offers a glance to people's perception toward AI. Because social perception always plays an important part as a forerunner for individuals' attitudes and behaviors (Fiske, 1993), it may contribute to predicting people's further attitudes and behaviors in their potential interaction with AI. Besides, our research could serve as a stepping-stone for understanding actual human–AI interactions. For example, individuals prefer utilitarian autonomous vehicles (sacrifice passengers for the greater good) in general and want people other than themselves to buy this type of vehicle (Bonnefon, Shariff, & Rahwan, 2016). A kind of stereotype – what we found in our research – might induce this effect in which people believe that the autonomous vehicle is an actor who should make utilitarian decisions in moral dilemmas. However, in line with our finding, as individuals positively interact with high-warmth partners (e.g., Samson & Zaleskiewicz, 2020), autonomous vehicles' potential low warmth may account for people's reluctance to purchase them. In addition, our findings suggest a potential explanation for people's disapproval of AI in moral decision-making. According to Bigman and Gray (2018), people are averse to AI's moral decision-making because AI is seen as lacking the capacity for experience. Due to the strong connection between "experience" and "warmth" (Gray and Wegner, 2012), people may oppose AI's involvement in moral issues because we are afraid that AI is likely to make utilitarian decisions. This is also consitent with previous findings that people generally hold negative attitudes toward utilitarian AI (Bonnefon et al., 2016).

Finally, our results can provide insights for AI's applications. Studies have found that people can be averse to AI in moral-related issues (Bigman & Gray, 2018; Longoni et al., 2019). As Studies 3 and 4 suggested, people seem not to be too strongly against AI for issues with less relevance to human lives (i.e., personal low-conflict dilemmas). In addition, in the example of autonomous vehicles, it might be easier for the public to accept them if those vehicles are used far away from human beings. As for AI's applications in other real-world settings, it could also be started in those scenarios without the possibility of potential harm.

8.2. Limitations and future direction

Nevertheless, our research is not without limitations. First, in line with previous researches (e.g., Chong, Zhang, Goucher-Lambert, Kotovsky, & Cagan, 2022; Lv, Yang, Qin, Cao, & Xu, 2022), we did not set a specific context or provide a very detailed discerption toward the so-called AI in our research, but only assign the AI with a particular role (i.e., decision-maker in moral dilemma). We suppose it is not a good idea to be that specific because there is still no AI in real moral dilemmas, and people will hold various assumptions toward AI. In other words, AIs seem like a kind of "Black Box" to most publics on matters in reality or in current research. However, people may also carry different perceptions toward AI regarding different contexts or descriptions, considering there are still a few researches that provided participants with a brief definition of AI (e.g., Gillath et al., 2021). For example, as Study 3 and 4 suggested, a context with less emotional evocation could result in fewer perceived differences between humans and AI. Future research may explore how AI's different moral-related applications could influence people's perception toward its behaviors. Secondly, the cross-cultural reliability of the main effect needs further examination, although Studies 3 and 4 suggested no cultural differences. Individuals in East Asia hold stronger moral-related values (e.g., conservatism) than their counterparts do in the West (Ros, Schwartz, & Surkiss, 1999). Furthermore, a recent study on moral dilemma of autonomous vehicles indicated that Canadians more than Koreans expected autonomous vehicles to be more utilitarian (Rhim, Lee, & Lee, 2020). Future research could investigate whether Asian people believe both AI and human beings make relatively fewer utilitarian choices, as compared to their Western counterparts.

In addition, we did not examine whether the perceived warmth in AI could be boosted through potential interventions. For instance, self-disclosure reduced individuals' tendency to blame computers for negative outcomes (Moon, 2003). Hence, more preinteraction with AI may help to some degree. Other researchers have indicated that presenting autonomous vehicles (similar to AI) with mentalistic (vs. mechanistic) terms could induce more trust and less blame to their behavior's negative results (Young & Monroe, 2019), which suggests that making AI more like humans may also work. Future research could test the effectiveness of potential interventions in boosting people's perceived warmth in AI, thus eliminating our reported main effects.

Besides, the current mediation model proposed in Study 2 should be interpreted cautiously because it does not allow inferences of causality. Even though the model in line with our theoretical illustration and two alternative models was ruled out, there could still be other alternative mediators in addition to warmth perceptions that might also work in the relation between the identity of decision-makers and people's prediction about its behavior in a moral dilemma. Future research may manipulate the mediator for a more convincing causal link if practical manipulations of warmth for our experimental scenario could be proposed. Nevertheless, as the warmth perception plays an important role in individuals' prediction about one entity's moral-dilemma behavior (Rom et al., 2017), we argue that warmth perception is still one plausible mediator.

Moreover, we only focused on people's predictions regarding AI in moral dilemmas . As Bigman and Gray (2018) found people averse AI's engagement in general moral issues, but whether people accept AI's engagement in moral dilemmas is unclear, and should be further explored, especially when an AI is not delegated to deal with related tasks because the ability to make moral judgments is considered unique to human beings (Opotow, 1990).

Lastly, we did not test how people respond to AI's decisions in moral dilemmas in which people's perceptions toward AI in a moral dilemma could induce further consequences. Because there are discrepancies between moral judgment and moral decision-making (Tassy, Oullier, Mancini, & Wicker, 2013), individuals' behavioral reactions and further attitudes toward AI's actions and their outcomes in a moral dilemma should be investigated in the future. For instance, we can ask participants if they want the particular AI (utilitarian or deontological) to be used in their daily lives, and what specific domains (e.g., help with driving or help with music selecting) with which they want AI to engage.

9. Conclusion

Our four studies provided converging support for the hypothesis that AI, relative to human beings, is perceived as being more likely to make utilitarian choices in moral dilemmas, and that this perceived warmth may account for that difference. We presented these findings using samples from different cultural backgrounds. A better understanding of how people perceive AI and predict their behaviors in moral dilemmas may guide a better regimentation for AI in the near future.

Ethical statement

The ethical approval (EA1912005) was received from the Human Research Ethics Committee (HREC) at the University of Hong Kong.

Disclosure statement

All data, materials, and codes for all studies can be found at https://osf.io/zdfnv/. Across all these studies, we report all measures, manipulations, and data exclusions, along with how we determine and justify our sample size.

Acknowledgements

We thank Prof. WANG Fang for her comments in the initial development of this project. We also thank Mr. SUN Tianxu for his comments on the research design.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104327.

References

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173–221). Mahwah, NJ: Lawrence Erlbaum Associates.

Bigman, Y., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bonnefon, J., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*, 1573–1576. https://doi.org/10.1126/science.aaf2654

Carmona-Perera, M., Caracuel, A., Pérez-García, M., & Verdejo-García, A. (2015). Brief moral decision-making questionnaire: A Rasch-derived short form of the Greene dilemmas. *Psychological Assessment, 27*, 424–432. https://doi.org/10.1037/pas0000049

Castelo, N., Bos, M., & Lehmann, D. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*, 809–825. https://doi.org/10.1177/0022243719851788

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior, 127*, Article 107018. https://doi.org/10.1016/j.chb.2021.107018

Cornelissen, G., Bashshur, M., Rode, J., & Le Menestrel, M. (2013). Rules or consequences? The role of ethical mind-sets in moral dynamics. *Psychological Science, 24*, 482–488. https://doi.org/10.1177/0956797612457376

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*, 273–292. https://doi.org/10.1177/1088868313495594

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*, 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082–1090. https://doi.org/10.1037/0022-3514.57.6.1082

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BF03193146

Fiske, S. (1993). Social cognition and social perception. *Annual Review of Psychology, 44*, 155–194. https://doi.org/10.1146/annurev.ps.44.020193.001103

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*, 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. In *Virtues and vices (p. Virtues and Vices, Chapter II)*. Oxford University Press. https://doi.org/10.1093/0199252866.003.0002.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior, 115*, Article 106607. https://doi.org/10.1016/j.chb.2020.106607

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics, 74*, 97–103. https://doi.org/10.1016/j.socec.2018.04.003

Gray, H., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*, 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*, 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108. https://doi.org/10.1126/science.1062872

Hayashi, Y., & Wakabayashi, K. (2018). Influence of robophobia on decision making in a court scenario. In *2018 ACM/IEEE international conference on human-robot interaction HRI '18* (pp. 121–122). https://doi.org/10.1145/3173386.3176988

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling. www.afhayes.com/public/process2012.pdf.

Hechler, S., & Kessler, T. (2018). On the difference between moral outrage and empathic anger: Anger about wrongful deeds or harmful consequences. *Journal of Experimental Social Psychology, 76*, 270–282. https://doi.org/10.1016/j.jesp.2018.03.005

Horowitz, M. (2016). The ethics and morality of robotic warfare: Assessing the debate over autonomous weapons. *Daedalus, 145*, 25–36. https://doi.org/10.1162/DAED_a_00409

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology, 89*, 899–913. https://doi.org/10.1037/0022-3514.89.6.899

Kant, I. (1785/1959). *Foundation of the metaphysics of morals (L. W. Beck, trans.)*. Bobbs-Merrill.

Kramer, M. F., Schaich Borg, J., Conitzer, V., & Sinnott-Armstrong, W. (2018). When do people want AI to make decisions?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 204–209). https://doi.org/10.1145/3278721.3278752

Liu, S. X., Shen, Q., & Hancock, J. (2021). Can a social robot be too warm or too competent? Older Chinese adults' perceptions of social robots and vulnerabilities. *Computers in Human Behavior, 125*, Article 106942. https://doi.org/10.1016/j.chb.2021.106942

Logg, J., Minson, J., & Moore, D. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research, 46*, 629–650. https://doi.org/10.1093/jcr/ucz013

Lv, X., Yang, Y., Qin, D., Cao, X., & Xu, H. (2022). Artificial intelligence service recovery: The role of empathic response in hospitality customers' continuous usage intention. *Computers in Human Behavior, 126*, Article 106993. https://doi.org/10.1016/j.chb.2021.106993

Mill, J. S. (1861/1998). In R. Crisp (Ed.), *Utilitarianism*. Oxford University Press.

Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology, 13*, 125–137. https://doi.org/10.1207/S15327663JCP13-1&2_11

Nass, C., Fogg, B., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human Computer Studies, 45*, 669–678. https://doi.org/10.1006/ijhc.1996.0073

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies, 43*, 223–239. https://doi.org/10.1006/ijhc.1995.1042

Opotow, S. (1990). Moral exclusion and injustice: An introduction. *Journal of Social Issues, 46*, 1–20. https://doi.org/10.1111/j.1540-4560.1990.tb00268.x

Rand, D. G. (2012). The promise of mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179. https://doi.org/10.1016/j.jtbi.2011.03.004

Rhim, J., Lee, G. B., & Lee, J. H. (2020). Human moral reasoning types in autonomous vehicle moral dilemma: A cross-cultural comparison of Korea and Canada. *Computers in Human Behavior, 102*, 39–56. https://doi.org/10.1016/j.chb.2019.08.010

Rom, S., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology, 74*, 24–37. https://doi.org/10.1016/j.jesp.2017.08.003

Rom, S., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58. https://doi.org/10.1016/j.jesp.2016.09.007

Ros, M., Schwartz, S. H., & Surkiss, S. (1999). Basic individual values, work values, and the meaning of work. *Applied Psychology, 48*, 49–71. https://doi.org/10.1111/j.1464-0597.1999.tb00048.x

Samson, K., & Zaleskiewicz, T. (2020). Social class and interpersonal trust: Partner's warmth, external threats and interpretations of trust betrayal. *European Journal of Social Psychology, 50*, 634–645. https://doi.org/10.1002/ejsp.2648

Shank, D., & Desanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Silver, D., Huang, A., Maddison, C. J., Guez, A., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*, 484–489. https://doi.org/10.1038/nature16961

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology, 4*, 250. https://doi.org/10.3389/fpsyg.2013.00250

Thomson, J. J. (1986). *Rights, restitution, and risk: Essays, in moral theory*. Harvard University Press. https://doi.org/10.2307/2215496

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*. https://doi.org/10.1016/j.jesp.2019.103870. in press.