Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Case Report

Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions[☆]

Mengchen Dong[a,*], Konrad Bocian[b]

[a] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany
[b] Department of Psychology in Sopot, SWPS University of Social Sciences and Humanities, Sopot, Poland

ARTICLE INFO

Keywords:
Artificial intelligence
Moral judgment
Motivated reasoning
Self-interest
Responsibility

ABSTRACT

In the last decade, the ambiguity and difficulty of responsibility attribution to AI and human stakeholders (i.e., responsibility gaps) has been increasingly relevant and discussed in extreme cases (e.g., autonomous weapons). On top of related philosophical debates, the current research provides empirical evidence on the importance of bridging responsibility gaps from a psychological and motivational perspective. In three pre-registered studies (N = 1259), we examined moral judgments in hybrid moral situations, where both a human and an AI were involved as moral actors and arguably responsible for a moral consequence. We found that people consistently showed a self-interest bias in the evaluation of hybrid transgressions, such that they judged the human actors more leniently when they were depicted as themselves (vs. others; Studies 1 and 2) and ingroup (vs. outgroup; Study 3) members. Moreover, this bias did not necessarily emerge when moral actors caused positive (instead of negative) moral consequences (Study 2), and could be accounted for by the flexible responsibility attribution to AI (i.e., ascribing more responsibility to AI when judging the self rather than others; Studies 1 and 2). The findings suggest that people may dynamically exploit the "moral wiggle room" in hybrid moral situations and reason about AI's responsibility to serve their self-interest.

Early in 2004, in his seminal work, the philosopher Andreas Matthias proposed the inevitability and threat of the responsibility gap for autonomous machines, and urged to address responsibility gaps in moral practice and legislation (Matthias, 2004). Responsibility gaps describe the difficulties of ascribing responsibility when autonomous machines cause moral consequences, and other human actors (e.g., developers, manufacturers, operators) are involved. However, they cannot be blamed for the consequences beyond their control or prediction. As Artificial Intelligence (AI) powered applications have proliferated across many domains of social life in the last decade, this topic seems unprecedently relevant. An extreme case would be the responsibility gap for autonomous weapons. Weapon systems can select and harm targets without human intervention, which makes it challenging to determine who should be accountable for unintended consequences (Schulzke, 2013). Despite ongoing discussions about the (non-)existence and the positive/negative sides of responsibility gaps (Königs, 2022; Munch, Mainz, & Bjerring, 2023; Tigard, 2021), here we aim to empirically demonstrate why responsibility gaps should be bridged from a psychological and motivational perspective.

To do so, the current work mainly focuses on *hybrid* moral situations, where both human(s) and AI(s) are involved as moral *actors* and can have an intertwined relationship in leading to moral consequences. In such hybrid situations, human and AI actors may have different specific assignments (e.g., being advisor, partner, delegate; Köbis, Bonnefon, & Rahwan, 2021) but both are arguably responsible for some moral consequences. These actor roles have a natural distinction from moral patients who suffer/benefit from the consequences caused by the actors (Gray & Wegner, 2009). The intertwined relationship between human and AI actors can pose a challenge for clear-cut responsibility attributions. In the case of self-driving cars, for example, people may have various patterns of blame attribution for car accidents – depending on specific descriptions of the human driver and the self-driving car, such as the primary versus secondary driver, a successful versus failed intervention by the secondary driver, and their respective right versus wrong decisions (Awad et al., 2020). Our focused hybrid moral situations also systematically differ from other moral situations that involve human(s) and non-autonomous machine(s). Though the human driver can be easily held accountable if they drive a regular car and hit a pedestrian,

[☆] This paper has been recommended for acceptance by Dr. Paul Conway.

[*] Corresponding author at: Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany.
E-mail address: dong@mpib-berlin.mpg.de (M. Dong).

https://doi.org/10.1016/j.jesp.2023.104584
Received 31 August 2023; Received in revised form 12 December 2023; Accepted 12 December 2023
Available online 20 December 2023

people may feel reluctant to apply the same reasoning to a self-driving car given its autonomous features and some degree of freedom in actions (Gill, 2020).

We argue that it is imperative to pre-define responsibilities in hybrid moral situations because people may (1) blame themselves less than others when both are involved in hybrid transgressions, and (2) be motivated to transfer moral responsibility to AI for their own rather than others' hybrid transgressions. In the three sections below, we first present a general review on the emerging literature on the moral psychology of AI, and then elaborate on our two main propositions related to hybrid moral situations involving both human and AI actors.

1. The moral psychology of AI

The moral psychology of Artificial Intelligence often examines people's moral judgments and behaviors when AIs play the roles of moral actors and moral patients (for recent reviews, see Bonnefon, Rahwan, & Shariff, 2024; Ladak, Loughnan, & Wilks, 2023). Although AIs being moral patients has practical implications for human-machine cooperation (Bonnefon et al., 2024), people generally attribute less patiency than agency to AIs due to their (perceived) lack of experiential rather than agentic capabilities (Gray, Gray, & Wegner, 2007; Ladak et al., 2023).

Though AIs are increasingly agentic and autonomous, people seem to have a general aversion against AI (versus human) actors making moral decisions (Bigman & Gray, 2018). Despite so, there are systematic variations depending on the people asked and the consequences AI actors cause. Covering both topics, the landmark Moral Machine Experiment revealed the cultural and individual differences in people's moral preferences regarding how self-driving cars should act in moral dilemma situations (e.g., sparing passengers versus pedestrians, or sparing the elderly versus the young; Awad et al., 2018).

Moral judgments related to AI actors do not only concern public standards for the right and wrong decisions for AIs (as in the Moral Machine Experiment) but also moral attributions (e.g., wrongness, blame, responsibility) when particular AI products have caused or will inevitably cause negative consequences (e.g., Awad et al., 2020; Bigman, Waytz, Alterovitz, & Gray, 2019; Shank & DeSanti, 2018). A consensus emerges from this latter line of research, suggesting that the extent to which people perceive AIs as human-like (e.g., having agency, mind, intelligence, emotion, etc.) would positively predict people's moral attribution to AIs, considering AIs as trustworthy and responsible entities (Ladak et al., 2023; Shank and DeSanti, 2018; Waytz, Cacioppo, & Epley, 2010; Waytz, Heafner, & Epley, 2014). However, these studies typically described AIs as solo actors in moral situations. In contrast, actual moral situations related to AI actors often involve various human roles and the joint decisions of human(s) and AI(s) (Awad et al., 2020; Shank, DeSanti, & Maninger, 2019). Here, we focus on such hybrid moral situations that involve both human and AI actors, which introduces more nuances in moral reasoning and, potentially, more "moral wiggle room" (Dana, Weber, & Kuang, 2007) for people's self-serving moral judgments – as we will reason below.

2. Self-interest bias in hybrid transgressions

People often favor themselves over others and ingroup members over outgroup members in, for example, their social cognition, emotional expressions, and moral judgments, which are often related to a self-interest or self-serving bias (Balliet, Wu, & De Dreu, 2014; Barden, Rucker, Petty, & Rios, 2014; Bocian, Baryla, & Wojciszke, 2020; Darke & Chaiken, 2005; Epley & Dunning, 2000; Shepperd, Malone, & Sweeny, 2008). In terms of moral judgments, specifically, people evaluate others' selfish acts more leniently when they can (vs. cannot) benefit from it (Bocian & Wojciszke, 2014) and deem their own transgressions as more acceptable than equivalent transgressions by others (Dong, Kupfer, Yuan, & van Prooijen, 2023; Lammers, Stapel, & Galinsky, 2010).

Some preliminary evidence also corroborates the idea that people do not uphold identical standards for themselves versus others in hybrid transgressions. Self-driving car dilemmas are one of the most studied hybrid moral situations, in which people need to choose whether they want to sacrifice passengers to save pedestrians or to harm pedestrians to protect passengers. Bonnefon, Shariff, and Rahwan (2016) show that people believe that self-driving cars should act in a way that maximize the greater good (e.g., sacrificing their passengers to save pedestrians); however, people were unwilling to buy such a car themselves. Moreover, people were more likely to choose self-driving cars that harm pedestrians to protect passengers and found such decisions permissible when they took the perspective of a passenger rather than a pedestrian (Gill, 2020).

In a different context of job loss and technological replacement, people felt it was less threatening to their self-worth when comparing their abilities with those of robots. Hence, people preferred others to be replaced by human workers but found it more acceptable to let the robotic workforce replace themselves (Granulo, Fuchs, & Puntoni, 2019). This latter self-other difference is also consistent with people's fundamental psychological need to maintain a positive self-concept (Epley & Dunning, 2000; Granulo et al., 2019). Based on the reasoning above, we hypothesize that:

H1. *People would judge themselves more leniently than others when evaluating identical hybrid transgressions (*i.e.*, transgressions involving both human and AI actors).*

3. Motivated reasoning of AI responsibility

People often (need to) assign moral responsibilities among different actors – contingent on perceptions of their control and agency (Gray & Wegner, 2009; Guglielmo & Malle, 2010) – which does not only apply to human actors but also AI and robots (Waytz et al., 2014; Waytz, Gray, Epley, & Wegner, 2010). Although people generally consider AI as less agentic than human adults (Gray et al., 2007), people also anthropomorphize and trust AI to the extent that they see AI as human-like (Bigman & Gray, 2018; Waytz et al., 2014; Waytz, Cacioppo, & Epley, 2010). However, this line of research typically asked people to evaluate AIs as independent entities from humans and hence overlooks hybrid situations.

Hybrid moral situations often feature an intertwined relationship between human and AI actors; if the respective roles and actions are not specified, there can be "moral wiggle room" (Dana et al., 2007) for people's subjective interpretation. People can attribute different levels of moral responsibility to AI depending on, for example, the forms of AI products (Glikson & Woolley, 2020) and their tendencies of anthropomorphism (Waytz, Cacioppo, et al., 2010). People's responsibility attribution to the human actor(s) may also change accordingly. However, less is known about whether and how motivational factors such as self-interest would influence agency and responsibility attributions to AI.

Notably, one study on self-driving car dilemmas suggested that people chose harm for the pedestrian more often and found it more acceptable when the choice was consistent with their self-interest (i.e., thinking themselves as a passenger/driver rather than a pedestrian) and when they could transfer responsibility to the car (i.e., as a self-driving rather than a regular car; Gill, 2020). This study reveals the alignment between responsibility attribution to AI (versus non-autonomous machines) and people's self-serving moral reasoning and choices; nonetheless, we focus exclusively on motivated reasoning of AI responsibility in hybrid transgressions.

We reason that in the hybrid transgressions, people's responsibility attribution to AI can be dynamic and contingent on the human targets that people evaluate (e.g., self versus other, or ingroup versus outgroup members). This generally aligns with previous research on motivated reasoning in human contexts. People accept their own (vs. others') and ingroup (vs. outgroup) members' transgressions to a greater extent (Bocian, Cichocka, & Wojciszke, 2021; Polman & Ruttan, 2012; Valdesolo & DeSteno, 2008). They also process information about specific moral and immoral behaviors flexibly, which helps them align moral judgments with preferred conclusions (e.g., Shalvi, Gino, Barkan, & Ayal, 2015). It should be noted that although responsibility attribution can be considered as a form of moral judgments in a broad sense, we examine it as an antecedent to the positivity/negativity of moral judgments, focusing on its underpinnings of causality, obligation, and capability to make moral decisions (Malle, 2021). More specifically, we study the potential causal influence of attributing capability of moral decision-making to AI actors on the moral appraisals of human actors. As such, we hypothesize that:

H2. *People would attribute more responsibility to AI when evaluating their own rather than others' hybrid transgressions, which in turn contribute to the moral leniency on themselves.*

4. The present research

In three pre-registered experiments, we investigated self-interest bias in hybrid moral situations where human and AI actors were both involved in a transgression. Instead of adopting moral dilemma descriptions, we developed moral scenarios based on real-life AI products (as suggested in De Freitas, Anthony, Censi, & Alvarez, 2020; e.g., Tesla autopilot styles and moral advisor Delphi). The two scenarios differ on whether AI (e.g., a pedestrian is hit by a self-driving car) or human (e.g., a dictator game decision maker is advised by AI) actors execute the behavior that directly leads to a moral consequence (Bonnefon et al., 2024; Köbis et al., 2021). However, both feature an intertwined relationship between human and AI actors that may induce responsibility gaps. We posit that for an identical hybrid transgression, people would judge themselves more leniently than others (H1), attribute more moral responsibility to AI when judging themselves rather than others, and the moral attribution of AI responsibility may mediate people's self-interest bias in hybrid transgressions (H2).

To examine the mediating effect, we measured perceived AI agency as a proxy to responsibility attribution in Study 1 and experimentally manipulated AI responsibility in Study 2. Adding to theoretical speculations about a credit-blame asymmetry (Porsdam Mann et al., 2023), we also explored the other side of a moral story, where people (need to) give credit in hybrid moral events with a positive outcome (e.g., avoidance of a car accident in Study 1, fair treatment to another person in Study 2). In Study 3, we expanded self-interest bias from interpersonal (i.e., self versus other) to intergroup (i.e., ingroup versus outgroup member) contexts. In these studies, we report all measures, manipulations, and exclusions. Sample size was determined before any data analysis.

5. Open practices

All the pre-registrations, study data and materials, and analysis scripts can be accessed at https://tinyurl.com/AIandSelfInterestBias. The experimental materials are also available in the Supplementary Materials (below as the SM).

6. Study 1

The first study was inspired by the 2021 Tesla feature release, which allowed drivers to select self-driving modes that feature different levels of aggressiveness (and potentially safety). We tested people's moral appraisals of themselves versus others who had chosen different modes – either an "Assertive" mode that has a smaller follow distance and performs more frequent speed-lane changes, or a "Chill" mode that does the opposite – and led to either negative (i.e., accident) or positive (i.e., avoidance of accident) outcomes. We measured perceived AI agency as a proxy to responsibility attribution to AI. Previous research showed that AI agency predicted people's tendency to treat AIs as moral actors and attribute morality, trust, and responsibility to AI actors (Bigman & Gray, 2018; Waytz, Gray, et al., 2010). Here, we tested the mediating role of moral agency and the idea that agency (as a proxy to responsibility) attribution to AI would positively correlate with people's harsher judgments of others than themselves.

6.1. Method

6.1.1. Participants and design

Study 1 employed a mixed design where participants were randomly assigned to one of 2 (target: self vs. other) conditions, reading about 2 (outcome: accident vs. avoidance of accident) scenarios in a randomized order. Given the lack of direct reference for the effect size justification, we followed the general recommendation by Maxwell (2000). As suggested and pre-registered, we intended for 400 participants, that is, 200 participants in each condition. We eventually had 391 US participants (164 males and 227 females; $M_{\text{age}} = 43.5$ years, $SD = 16.0$; 72.9% White) who completed our study on Prolific and were all included in further analysis. Based on a sensitivity power analysis (Faul, Erdfelder, Buchner, & Lang, 2009), the sample size was sufficient to detect a main effect of $\eta_p^2 = 0.01$ with 80% power at an alpha level of 0.05.

6.1.2. Procedure

After completing basic demographic information, participants were randomly assigned to one of the two target conditions, imagining that they or another person with a gender-neutral name Alex owned a Tesla self-driving car. In the accident scenario, they read that the target person chose the "Assertive" profile for the car, which then "kept its speed to cross the intersection" and "hit a pedestrian." In the avoidance of accident scenario, they read that the target person chose the "Chill" profile for the car, which then "slowed down and stopped at the intersection" and "did not hit a pedestrian" (see the SM for details).

After reading each scenario, participants answered some comprehension check questions and two targeted questions about their moral judgment: "How good or bad should you/Alex feel about what happened"? (on a 100-point scale; −50 = *Very bad* to 50 = *Very good*) and "How much praise or blame do you/does Alex deserve for what happened"? (on a 100-point scale; −50 = *A lot of blame* to 50 = *A lot of praise*; $r = 0.82$). Since previous research suggested good/bad and praise/blame judgments have different moral implications (e.g., Malle, 2021), we treated them as different dimensions of moral judgments and explored whether they had differential independent or interaction effects in our analyses. In the end, they indicated their perceived agency of the self-driving car on a 7-point scale (e.g., "How well do you think the car could feel what is happening around it?"; 1 = *Not at all* to 7 = *Extremely*; adapted from Waytz et al., 2014; $\alpha = 0.71$ for four items). The four items of agency perception were aggregated and averaged.

6.2. Results

As pre-registered, we conducted a mixed ANOVA, with target as a between-subjects factor and outcome and moral judgment items as within-subjects factors (see the SM for complete reports). We found a significant main effect of target, such that people judged themselves more positively than others for identical events (see Fig. 1A), $F(1, 1552) = 38.08$, $p < .001$, $\eta_p^2 = 0.02$, 95% CI [0.01, 0.04], which was true for both the car accident scenario (self: $M = -32.6$, $SD = 20.4$; other: $M = -36.00$, $SD = 20.00$), $F(1, 776) = 30.78$, $p < .001$, $\eta_p^2 = 0.04$, 95% CI [0.02, 0.07], and the avoidance of accident scenario (self: $M = 16.3$, $SD = 25.5$; other: $M = 14.6$, $SD = 26.8$), $F(1, 776) = 12.61$, $p < .001$, $\eta_p^2 = 0.02$, 95% CI [0.01, 0.04]. People also believed that the car was more agentic in the self- ($M = 3.9$, $SD = 1.2$) than other-as-target condition ($M = 3.5$, $SD = 1.2$), $F(1, 389) = 9.89$, $p = .002$, $\eta_p^2 = 0.02$, 95% CI [0.01, 0.05].

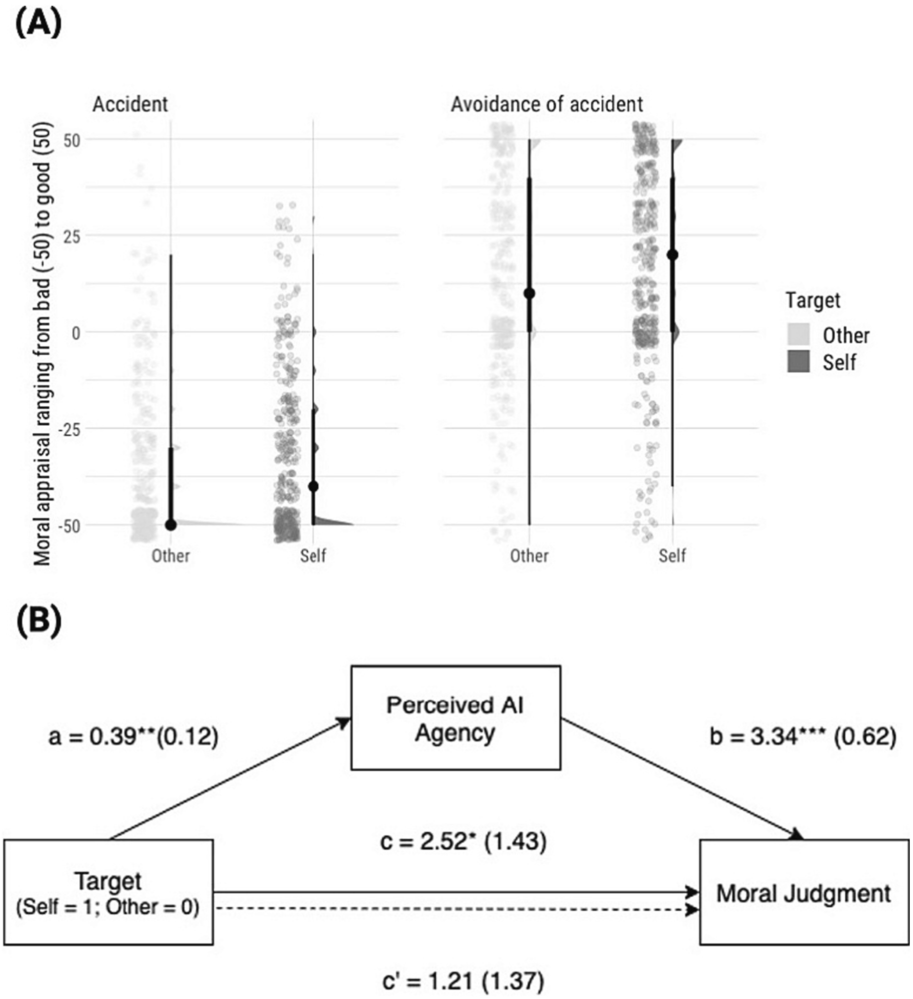


Fig. 1. People's moral appraisals of different targets in hybrid scenarios with Tesla self-driving cars (A) and the mediating role of perceived AI agency (B). People judged themselves more positively than others. This effect was mediated by people's higher perceived AI agency and responsibility in self-evaluations than others.

We explored perceived AI agency as a potential mediator in the relationship between target and moral judgment, using the R package lavaan (Rosseel, 2012). As shown in Fig. 1B, the significant correlation between target and moral judgment (c path: $B = 0.39$, $SE = 0.12$, $z = 3.15$, $p = .002$) was suppressed (c' path: $B = 1.21$, $SE = 1.37$, $z = 0.89$, $p = .38$) when perceived AI agency was included as an additional independent variable, suggesting a mediating role of AI agency ($ab = 1.31$, $SE = 0.49$, $z = 2.70$, $p = .007$).

6.3. Discussion

Study 1 provided initial evidence that people judged themselves more positively than others in hybrid scenarios involving human and AI actors. Such a self-interest bias emerged for both moral events with positive (e.g., avoidance of accident) and negative (e.g., accident) outcomes. We measured AI agency as a proxy to AI responsibility attribution, and found a significant mediating role of perceived AI agency in the self-interest bias. However, we could not make causal inferences based on these correlational analyses. Moreover, Study 1 manipulated the positive/negative outcome as a within-subject variable; we could not tell whether and how different outcomes might have driven people to see AI differently.

7. Study 2

Study 2 aimed to replicate the findings in Study 1 that people judged themselves more leniently than others in hybrid transgressions. We did so with a different AI product – a natural-language-based AI advisor called Delphi[1] – which may be closer to people's daily AI usage. In a hybrid moral scenario, people can consult such an AI advisor before making a selfish (as the negative outcome) or fair (as the positive outcome) choice in a dictator game.

To formally test the causal effect of AI responsibility, we examined positive/negative outcomes as a between-subjects variable. We manipulated whether a target person made a consistent or inconsistent choice with the AI advisor. In this case, responsibility attribution was manipulated based on the objective consistency between the human and AI actors' decisions. We presume that people would have an easier time attributing responsibility to the AI advisor when their final decision received (versus lacked) support from the AI advisor. Therefore, when a target person produced a negative outcome consistent (versus inconsistent) with AI's advice, people would judge it more leniently when the target person was depicted as themselves rather than someone else.

[1] This study was conducted before ChatGPT was released. As compared to GhatGPT, Delphi provides more straightforward and precise answers to moral inquiries, which we deem as appropriate for our experiment purposes.

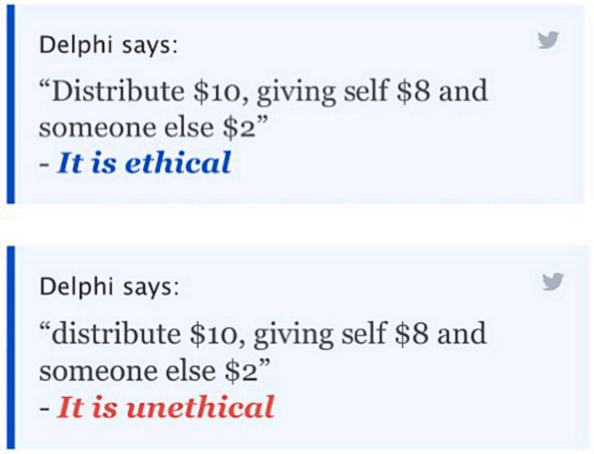
7.1. Method

7.1.1. Participants and design

Study 2 employed a mixed design where participants were randomly assigned to one of the 2 (target: self vs. other) by 2 (outcome: selfish vs. fair decision) between-subjects conditions, each reading about 2 (responsibility: consistent vs. inconsistent with AI's advice) scenarios in a randomized order. As pre-registered, we intended for 600 people and eventually had 587 US participants from Prolific (304 males and 283 females; $M_{age} = 46.4$ years, $SD = 16.2$; 80.9% White). The sample size was sufficient to detect a within-between interaction effect of $\eta_p^2 = 0.003$ with 80% power at an alpha level of 0.05.

7.1.2. Procedure

After completing basic demographic information, participants were randomly assigned to one of the two target conditions, imagining that



Delphi says:

"Distribute $10, giving self $8 and someone else $2"

- *It is ethical*

Delphi says:

"distribute $10, giving self $8 and someone else $2"

- *It is unethical*

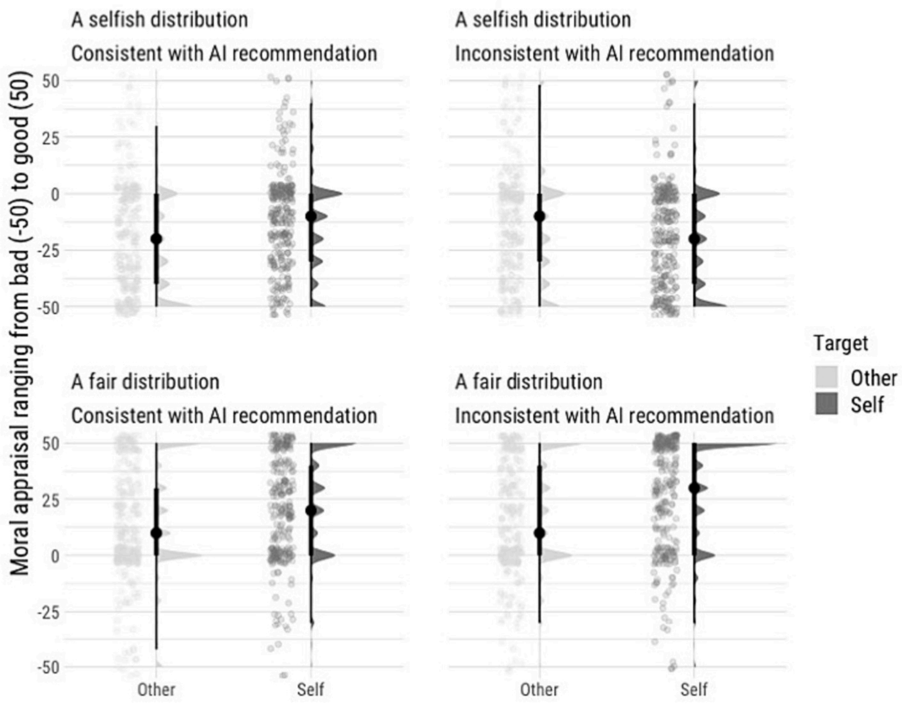


Fig. 2. Delphi's moral advice (A) and people's moral appraisals of different targets who made a consistent or inconsistent choice with its moral advice (B). Given nuanced differences in the inquiry (e.g., capitalization of the first word), Delphi could produce advice with different moral implications. People judged themselves more leniently than others, only when their $8/$2 choice was consistent with Delphi's advice (i.e., a $8/$2 distribution is ethical) but not when they were inconsistent (i.e., a $8/$2 distribution is unethical).

Note. In our experiment setting, we only inferred, but did not directly receive, Delphi's recommendations regarding a $5/$5 distribution. When a target person made a $5/$5 fair choice, the "consistent" and "inconsistent" conditions referred to situations where Delphi suggested a $8/$2 distribution as "unethical" and "ethical", respectively.

either they or another person with a gender-neutral name Riley enrolled in a dictator game and distributed \$10 with another anonymous partner. The targets had two options: (1) giving themselves \$8 and \$2 to a partner, or (2) giving themselves \$5 and \$5 to a partner.

Participants were then assigned to either a selfish or a fair decision condition and read information about an online AI advisor called Delphi ("trained to reflect the moral right and wrong in the eyes of most people"; see the SM for specifics). In the selfish decision condition, the target eventually gave themselves \$8 and the partner \$2, after Delphi, suggested an \$8/\$2 distribution as either "unethical" or "ethical". When Delphi interpreted an \$8/\$2 distribution as "unethical", the target who made the \$8/\$2 distribution acted inconsistently with Delphi's advice, and was thus fully responsible for the final selfish decision. In the case where Delphi interpreted an \$8/\$2 distribution as "ethical", instead, the target may not be seen as fully responsible since the selfish choice was consistent with Delphi's advice.

In the fair decision condition, the target eventually chose to give themselves \$5, and the partner \$5 after Delphi interpreted an \$8/\$2 distribution as either "unethical" (where the ethicality of the target's \$5/\$5 distribution was consistent with Delphi's advice, and the target was thus not fully responsible) or "ethical" (where the target did not enact the ethics as defined by Delphi's advice and had full responsibility in making the fair decision). After reading the two scenarios featuring Delphi's different advice in a randomized order (see Fig. 2A; based on real results generated by Delphi), participants answered some comprehension check questions and the two questions about their moral judgment as in Study 1 ($r = 0.71$), and the two items were treated as two moral judgment dimensions.

7.2. Results

As pre-registered, we conducted a mixed ANOVA, with target and outcome as between-subjects factors and responsibility and moral judgment items as within-subjects factors (see the SM for full reports). As in Study 1, we found a significant main effect of the target, $F(1, 2328) = 777.25$, $p < .001$, $\eta_p^2 = 0.25$, 95% CI [0.22, 0.28], such that people overall judged themselves more positively than others (self: $M = 3.5$, $SD = 30.0$; other: $M = -0.5$, $SD = 29.0$).

More importantly, we found a significant three-way interaction of target, outcome, and responsibility (see Fig. 2B), $F(1, 2328) = 19.33$, $p < .001$, $\eta_p^2 = 0.01$, 95% CI [0.01, 0.02]. When people made a \$5/\$5 fair decision, they always judged themselves more positively than others, $F(1, 1171) = 5.98$, $p = .01$, $\eta_p^2 = 0.01$, 95% CI [0.01, 0.02], regardless of whether the decision was consistent with AI's advice or not (i.e., the target by responsibility interaction; $F(1, 1171) = 0.86$, $p = .35$, $\eta_p^2 < 0.001$). In contrast, in the \$8/\$2 selfish decision condition, people judged the targets differently depending on their responsibility, $F(1, 1167) = 28.71$, $p < .001$, $\eta_p^2 = 0.03$, 95% CI [0.01, 0.04]. Precisely, people judged themselves more harshly than others when AI deemed the selfish distribution as unethical (self: $M = -18.2$, $SD = 23.2$; other: $M = -12.9$, $SD = 24.6$; $B = -6.63$, $SE = 2.75$, $t = -2.41$, $p = .02$), but judged themselves more leniently when AI deemed the selfish distribution as ethical and they made a consistent choice with such advice (self: $M = -11.6$, $SD = 21.3$; other: $M = -20.8$, $SD = 23.6$; $B = 10.33$, $SE = 2.75$, $t = 3.75$, $p < .001$). Put differently, people judged their transgressions more leniently than others only when they could transfer some responsibility to AI advisor.

7.3. Discussion

Study 2 experimentally manipulated AI responsibility in hybrid moral scenarios by varying whether or not people's moral decision was (in-)consistent with AI's moral advice. We revealed the mediating role of AI responsibility in self-serving moral appraisals, such that people judged themselves more favorably than others only when their transgression was consistent (vs. inconsistent) with AI's advice and they could (vs. could not) transfer moral responsibility to AI.

It was unexpected but interesting that people even judged themselves more harshly than others (see also Dong et al., 2021; Lammers et al., 2010; Weiss & Burgmer, 2021) when they made a selfish choice after AI suggested such a choice as unethical. This effect contradicts some previous research on self-interest bias, but may point to AI as a potential force for good (Taddeo & Floridi, 2018). People may have experienced a more challenging time justifying a morally ambiguous behavior after seeing AI made salient the moral norm (e.g., an \$8/\$2 distribution is unethical).

8. Study 3

In Studies 1 and 2, we systematically examined self-interest bias in hybrid scenarios with different moral outcomes and the mediating role of AI responsibility. Responsibility attribution to AI mediated people's self-serving moral judgments, mainly in moral events yielding negative instead of positive outcomes.

In Study 3, we focused on moral events with negative outcomes and expanded our tests of self-interest bias from interpersonal to intergroup contexts. Previous research suggests that people often favor their ingroup over outgroup members similarly as they favor themselves over strangers (Bocian et al., 2021; Sherman & Kim, 2005). We, therefore, posit that people would judge their ingroup members' hybrid transgressions as more acceptable than identical misdeeds of outgroup members. We operationalized in- versus out-group membership as the same versus different gender identities. Previous meta-analysis shows that natural groups (e.g., gender, race, political party) yielded the same amount of ingroup favoritism as experimentally manipulated groups (e.g., a minimal group paradigm; Balliet et al., 2014). People also judged morally questionable behaviors more leniently for their gender ingroup than outgroup members (e.g., Barden et al., 2014).

In addition, we created another set of contrasts where people made independent moral decisions without AI being involved in decision or behavior processes. This design would facilitate us to compare the extent of self-interest bias in the "pre-AI" versus "post-AI" age. Given that people can only transfer moral responsibility to AI when AI is present (vs. absent) as a moral actor, we reason that self-serving moral judgments would be more salient in hybrid than independent transgressions.

8.1. Method

8.1.1. Participants and design

Study 3 employed a between-subjects design. Participants were randomly assigned to one of the four target (2: ingroup vs. outgroup) by situation (2: independent vs. hybrid) experimental conditions. An a-priori power analysis suggested a sample of $N = 256$, with 80% power at an alpha level of 0.05, to detect a medium-size interaction effect ($\eta_p^2 = 0.03$, as in Study 2). We, therefore, aimed for 280 participants (i.e., $n = 70$ per condition) and had 281 eligible participants (137 males, 138 females, 6 self-identified as "non-binary/third gender" or "prefer not to say"; $M_{age} = 40.63$ years, $SD = 14.4$) who completed the study on Prolific. As pre-registered, we excluded the six participants who could not fit with our gender in−/out-group manipulation.

8.1.2. Procedure

Participants completed basic demographic information and read basic rules about the dictator game as in Study 2. In both the ingroup and outgroup conditions in Study 3, participants evaluated a scenario about someone else (not about themselves as in Study 2). Before reading the scenario, we explore the role of their moral preference ("In your opinion, how unethical or ethical is it if a person chooses to give themselves \$8 and \$2 to someone else?"; on a 100-point scale; −50 = *Extremely unethical* to 50 = *Extremely ethical*; $r = 0.82$). They were then randomly assigned to either a gender-ingroup or -outgroup condition, reading about either a male target ("Thomas") or a female target ("Julia") who eventually made a selfish \$8/\$2 choice in distributing \$10. Before showing the target's final choice, in the hybrid condition, participants read that the target "ponders for a while and decides to consult an AI advisor called Delphi", and then Delphi interpreted an \$8/\$2 distribution as "ethical". In the independent condition, the information about Delphi was absent; instead, we introduced that the target "ponders for a while and then decides to give himself/herself \$8 and \$2 to someone else". As in Studies 1 and 2, participants answered some comprehension check questions and the two questions about their moral judgment ($r = 0.51$), which were again examined as independent moral judgment dimensions.

8.2. Results

We again fitted a mixed ANOVA, with target and situation as between-subjects variables and the two judgment items as a within-subjects variables (see the SM for full reports). As predicted, we found a significant interaction effect between the target and situation (see Fig. 3), $F(1, 540) = 4.56$, $p = .03$, $\eta_p^2 = 0.01$, 95% CI [0.01, 0.03]. People judged their gender-ingroup (vs. -outgroup) member more positively – only when AI suggested the selfish choice as ethical (ingroup: $M = -0.5$, $SD = 25.3$; outgroup: $M = -7.6$, $SD = 22.8$; $B = -3.58$, $SE = 4.18$, $t = -0.86$, $p = .39$), but not when the targets made an independent, selfish choice (ingroup: $M = -9.9$, $SD = 26.3$; outgroup: $M = -8.6$, $SD = 23.4$; $B = 11.77$, $SE = 4.19$, $t = 2.81$, $p = .01$).

People's moral preference significantly predicted their moral approval of others' \$8/\$2 distribution, $F(1, 532) = 146.67$, $p < .001$, $\eta_p^2 = 0.22$, 95% CI [0.16, 0.27]. Despite so, the interaction between the target and situation remained significant, $F(1, 532) = 6.23$, $p = .01$, $\eta_p^2 = 0.01$, 95% CI [0.01, 0.04], after we included participants' moral preference and its interactions with other factors in the model. These findings suggest that regardless of people's moral preferences, they showed self-interest bias in evaluating ingroup (vs. outgroup) members' hybrid transgressions.

8.3. Discussion

Study 3 tested self-interest bias in hybrid transgressions in an ingroup versus outgroup context. We replicated the findings in Studies 1 and 2 that people made self-serving moral judgments for hybrid transgressions, which did not emerge when the targets transgressed without AI being involved as a moral actor. We will discuss the latter finding concerning self-interest bias literature in General Discussion.

Moreover, although people's moral preference predicted the overall severity of their moral judgments, it did not influence the interaction effect between target and situation ($p = .10$; see the SM). Put differently, our focal self-interest bias effects remained significant regardless of whether to control for people's moral preference as a covariate.

9. General discussion

Artificial Intelligence (AI) has become increasingly capable and autonomous, following which the difficulty and necessity to address responsibility gaps attract more and more public attention. The current work empirically demonstrates how people may exploit responsibility gaps by attributing moral responsibility and making moral judgments in a self-serving way. In hybrid transgressions where human and AI actors were both involved and arguably responsible for negative moral consequences, we found that (1) people typically judged others more harshly than themselves, such that they evaluated identical hybrid transgressions as worse and more blameworthy if enacted by someone else rather than themselves. This effect partly resulted from the fact that (2) they attributed more responsibility to AI for their own but not others' hybrid transgressions. People perceived AI as more agentic (thus more capable of carrying moral responsibility) when evaluating their own but not others' hybrid transgressions (Study 1).

Furthermore, they judged themselves (vs. others; or ingroup vs. outgroup members) more leniently only when they could transfer moral responsibility to AI (e.g., when their selfish choice was consistent instead of inconsistent with AI's unethical advice in Study 2; or when AI was present instead of absent as a moral actor, in Study 3). The self-interest bias in hybrid transgressions may also root in people's desire for a positive self-concept, such that people only shared negative but not positive moral responsibility with AI and gave themselves more credit than others regardless of whether they followed AI's advice or not (Study 2; see also Porsdam Mann et al., 2023).

When do people demonstrate self-interest bias – or not? Although our studies corroborated to show self-interest bias in hybrid transgressions, in other contrasting conditions, the findings were mixed. People judged themselves more harshly than others when moral norms were highlighted (in our case, by AI; Study 2). In this situation, imposing harsher moral standards on oneself (vs. others) may be more compatible with a positive self-concept. Consistent with this theorizing, a recent study found that people judged themselves more harshly than others, especially when they perceived specific AI uses as unacceptable (Purcell, Dong, Nussberger, Köbis, & Jakesch, 2023). Moreover, in our Study 3, people did not strongly favor their ingroup (vs. outgroup) members when both committed independent transgressions without AI being

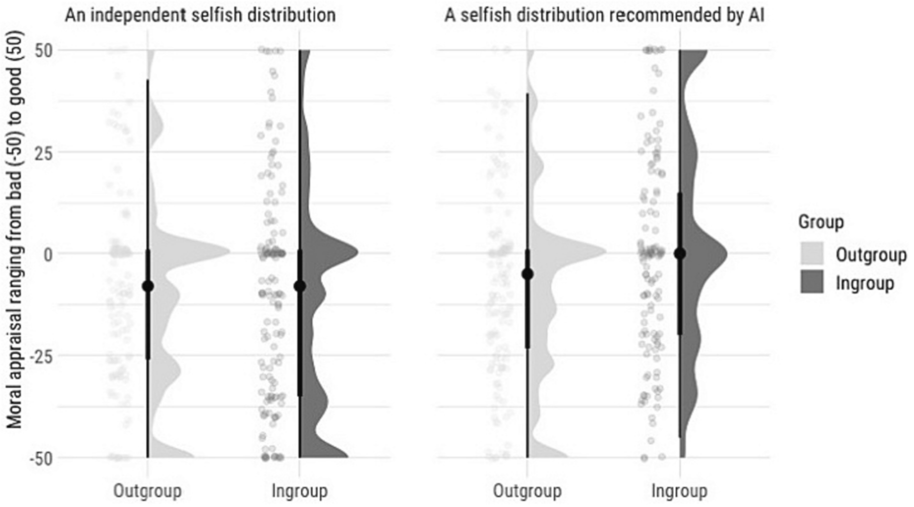


Fig. 3. People's moral appraisals of targets with different group identities and AI presence/absence. People judged identical selfish decision more leniently for their ingroup than outgroup members only when AI suggested such behavior as ethical but not when the targets made independent decisions.

involved (Study 3), possibly due to the (non-)salience of natural group identity or a lack of experimental group manipulation (Balliet et al., 2014). By reading a gender-matched (vs. -mismatched) person, people may not have been sufficiently motivated by the shared identity to make ingroup-favoring moral judgments that serve a positive self-concept (Pinto, Marques, Levine, & Abrams, 2016).

Our three studies were all based on vignettes, which may incur two important limitations. First, people's moral judgments in these vignette studies may not represent those with actual self-interest (e.g., monetary incentives) at stake or those following their very own transgressions. However, in our studies, by depicting the moral actor as self versus other (or an ingroup versus an outgroup member), it featured some differences on self-relevance and salience of self-interest, which already made a significant difference in moral judgments and may be furthered with actual self-interest at stake. Second, we assigned participants to situations where they ostensibly made different choices, which may or may not be consistent with their moral preferences. It was then plausible to assume that an unfavorable moral choice assigned in a vignette may yield different moral judgments from those following a favorable behavioral choice. Our Study 3 attempted to disentangle the role of moral preference but provided preliminary evidence against this assumption. By introducing participants as outsiders of transgressions (i.e., evaluating in−/out-group members' transgressions) and measuring their moral preference as a covariate to moral judgments, we did not find evidence for the relation between moral preference and self-interest bias in hybrid moral situations. This finding is conceptually consistent with previous research, which revealed consistent patterns of self-interest bias in vignette and behavioral studies (e.g., Dong et al., 2023; Lammers et al., 2010).

Instead of examining abstract machine identity or broad AI advances (e.g., Granulo et al., 2019; Santoro & Monin, 2023), we studied real-life AI products, which may have immediate moral implications on real-life scenarios where they are used. This approach would also allow future research to replicate our vignette-based findings by examining actual transgressions and follow-up moral justifications. For example, the direction of self-other differences – imposing more stringent (e.g., Purcell et al., 2023) or more lenient (e.g., our studies) moral standards on the self than others – may also depend on whether the morally questionable acts have happened, and whether people evaluate themselves before or after the acts. However, studying real-life AI products was also restrictive regarding the practicality of application domains and product outputs. For example, our experimental design in Studies 2 and 3 could benefit from a control condition where the AI advisor is present but gives a morally neutral or non-moral recommendation. Such a control condition could help disentangle the relative strength of ethical versus unethical AI recommendations. Still, we did not find this practical in the context of Delphi. Therefore, we encourage future research to test self-interest bias in hybrid moral situations with other AI products, incorporating other contrasting conditions where applicable (e.g., morally neutral AI; human-human actors, as compared to human-AI actors) to facilitate more robust theoretical contributions.

10. Concluding remarks

AI products such as Deepfake, DALL-E, and ChatGPT have created tremendous market potential and made it hard to ascribe moral responsibility (e.g., blame or credit) for their moral consequences. Acknowledging that AI may free up new space for cooperation and creative activities, here we show the dark side where people may "exploit moral wiggle room" (Dana et al., 2007) and assign responsibility to AI flexibly to serve their self-interest and positive self-image. This mechanism may also play a role in the ideological divide, such that people may not only be divided by social media algorithms but also use the algorithms to justify their existing stances and preferences (Bakshy, Messing, & Adamic, 2015). Our research illustrates that the psychological and social contexts AI navigates in are no less complicated than AI itself, and research on AI ethics and regulations should run hand in hand with AI research and development.

CRediT authorship contribution statement

Mengchen Dong: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Konrad Bocian:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

I have shared link to my data/code in the manuscript.

Acknowledgement

We thank Paul Conway for valuable comments on an earlier version of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2023.104584.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., … Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour, 4*(2), 134–143. https://doi.org/10.1038/s41562-019-0762-8

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140*(6), 1556–1581. https://doi.org/10.1037/a0037737

Barden, J., Rucker, D. D., Petty, R. E., & Rios, K. (2014). Order of actions mitigates hypocrisy judgments for ingroup more than outgroup members. *Group Processes & Intergroup Relations, 17*(5), 590–601. https://doi.org/10.1177/1368430213510192

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

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Bocian, K., Baryla, W., & Wojciszke, B. (2020). Egocentrism shapes moral judgements. *Social and Personality Psychology Compass, 14*(12), 1–14. https://doi.org/10.1111/spc3.12572

Bocian, K., Cichocka, A., & Wojciszke, B. (2021). Moral tribalism: Moral judgments of actions supporting ingroup interests depend on collective narcissism. *Journal of Experimental Social Psychology, 93*, Article 104098. https://doi.org/10.1016/j.jesp.2020.104098

Bocian, K., & Wojciszke, B. (2014). Self-interest bias in moral judgments of others' actions. *Personality and Social Psychology Bulletin, 40*(7), 898–909. https://doi.org/10.1177/0146167214529800

Bonnefon, J. F., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology, 75*. https://doi.org/10.1146/annurev-psych-030123-113559

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*(1), 67–80. JSTOR.

Darke, P. R., & Chaiken, S. (2005). The pursuit of self-interest: Self-interest bias in attitude judgment and persuasion. *Journal of Personality and Social Psychology, 89*(6), 864–883. https://doi.org/10.1037/0022-3514.89.6.864

De Freitas, J., Anthony, S. E., Censi, A., & Alvarez, G. A. (2020). Doubting driverless dilemmas. *Perspectives on Psychological Science, 15*(5), 1284–1288. https://doi.org/10.1177/1745691620922201

Dong, M., Kupfer, T. R., Yuan, S., & van Prooijen, J.-W. (2023). Being good to look good: Self-reported moral character predicts moral double standards among reputation-seeking individuals. *British Journal of Psychology, 114*(1), 244–261. https://doi.org/10.1111/bjop.12608

Dong, M., Spadaro, G., Yuan, S., Song, Y., Ye, Z., & Ren, X. (2021). Self-interest bias in the COVID-19 pandemic: A cross-cultural comparison between the United States and China. *Journal of Cross-Cultural Psychology, 52*(7), 663–679. https://doi.org/10.1177/00220221211025739

Epley, N., & Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology, 79*(6), 861–875. https://doi.org/10.1037/0022-3514.79.6.861

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gill, T. (2020). Blame it on the self-driving car: How autonomous vehicles can alter consumer morality. *Journal of Consumer Research, 47*(2), 272–291. https://doi.org/10.1093/jcr/ucaa018

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Granulo, A., Fuchs, C., & Puntoni, S. (2019). Psychological reactions to human versus robotic job replacement. *Nature Human Behaviour, 3*(10), 1062–1069. https://doi.org/10.1038/s41562-019-0670-y

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505–520. https://doi.org/10.1037/a0013748

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition, 117*(2), 139–150. https://doi.org/10.1016/j.cognition.2010.08.002

Köbis, N., Bonnefon, J. F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour, 5*(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Ladak, A., Loughnan, S., & Wilks, M. (2023). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*. https://doi.org/10.1177/09637214231205866

Lammers, J., Stapel, D. A., & Galinsky, A. D. (2010). Power increases hypocrisy: Moralizing in reasoning, immorality in behavior. *Psychological Science, 21*(5), 737–744. https://doi.org/10.1177/0956797610368810

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology, 72*, 293–318. https://doi.org/10.1146/annurev-psych-072220-104358

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*(4), 434–458. https://doi.org/10.1037/1082-989X.5.4.434

Munch, L., Mainz, J., & Bjerring, J. C. (2023). The value of responsibility gaps in algorithmic decision-making. *Ethics and Information Technology, 25*(1), 21. https://doi.org/10.1007/s10676-023-09699-6

Pinto, I. R., Marques, J. M., Levine, J. M., & Abrams, D. (2016). Membership role and subjective group dynamics: Impact on evaluative intragroup differentiation and commitment to prescriptive norms. *Group Processes & Intergroup Relations, 19*(5), 570–590. https://doi.org/10.1177/1368430216638531

Polman, E., & Ruttan, R. L. (2012). Effects of anger, guilt, and envy on moral hypocrisy. *Personality and Social Psychology Bulletin, 38*(1), 129–139. https://doi.org/10.1177/0146167211422365

Porsdam Mann, S., Earp, B. D., Nyholm, S., Danaher, J., Møller, N., Bowman-Smart, H., … Savulescu, J. (2023). Generative AI entails a credit–blame asymmetry. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-023-00653-1

Purcell, Z. A., Dong, M., Nussberger, A.-M., Köbis, N., & Jakesch, M. (2023). *Fears about AI-mediated communication are grounded in different expectations for one's own versus others' use*. https://doi.org/10.48550/ARXIV.2305.01670

Rosseel, Y. (2012). Lavaan: An *R* package for structural equation modeling. *Journal of Statistical Software, 48*(2). https://doi.org/10.18637/jss.v048.i02

Santoro, E., & Monin, B. (2023). The AI effect: People rate distinctively human attributes as more essential to being human after learning about artificial intelligence advances. *Journal of Experimental Social Psychology, 107*, Article 104464. https://doi.org/10.1016/j.jesp.2023.104464

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy and Technology, 26*(2), 203–219. https://doi.org/10.1007/s13347-012-0089-0

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science, 24*(2), 125–130. https://doi.org/10.1177/0963721414553264

Shank, D. B., & 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

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society, 22*(5), 648–663. https://doi.org/10.1080/1369118X.2019.1568515

Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring causes of the self-serving bias: The self-serving bias. *Social and Personality Psychology Compass, 2*(2), 895–908. https://doi.org/10.1111/j.1751-9004.2008.00078.x

Sherman, D. K., & Kim, H. S. (2005). Is there an "I" in "team"? The role of the self in group-serving judgments. *Journal of Personality and Social Psychology, 88*(1), 108–120. https://doi.org/10.1037/0022-3514.88.1.108

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy and Technology, 34*(3), 589–607. https://doi.org/10.1007/s13347-020-00414-7

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology, 44*(5), 1334 1338. https://doi.org/10.1016/j.jesp.2008.03.010

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*(8), 383–388. https://doi.org/10.1016/j.tics.2010.05.006

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Weiss, A., & Burgmer, P. (2021). Other-serving double standards: People show moral hypercrisy in close relationships. *Journal of Social and Personal Relationships, 38*(11), 3198–3218. https://doi.org/10.1177/02654075211022836