I Need Your Advice... Human Perceptions of Robot Moral Advising Behaviors

Nichole D. Starr
nstarr@mines.edu
Colorado School of Mines
Golden, Colorado

Bertram Malle
bfmalle@mines.edu
Brown University
Providence, Rhode Island

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, Colorado

ABSTRACT

Due to their unique persuasive power, language-capable robots must be able to both act in line with human moral norms and clearly and appropriately communicate those norms. These requirements are complicated by the possibility that humans may ascribe blame differently to humans and robots. In this work, we explore how robots should communicate in moral advising scenarios, in which the norms they are expected to follow (in a moral dilemma scenario) may be different from those their advisees are expected to follow. Our results suggest that, in fact, both humans and robots are judged more positively when they provide the advice that favors the common good over an individual's life. These results raise critical new questions regarding people's moral responses to robots and the design of autonomous moral agents.

KEYWORDS

Robot Ethics, Human-Robot Interaction, Moral Psychology

1 INTRODUCTION

Research in the HRI literature has consistently demonstrated that language-capable robots have significant persuasive power and are able to influence, persuade, and coerce humans in a variety of ways [3, 4, 11, 24, 25, 27, 33]. Moreover, there is evidence that this persuasive capacity has the potential for long-term impact, with robots exerting influence over their interactants' social and moral norms [6, 7, 14, 29]. This persuasive impact may not only influence those interactants' long-term social and moral behaviors, but also what behaviors those interactants choose to condone or sanction in others; in turn, this can lead to potential "ripple effects" across robots' social and moral ecosystems. As argued in previous work, this imposes unique moral responsibility on robots [8], and it suggests that if we are to develop and deploy language-capable social robots, they must have the requisite *moral competence* to avoid negatively impacting their social and moral ecosystem.

Malle and Scheutz propose four key requirements for robotic moral competence: [15, 18]: (1) a moral core (a system of moral norms and a moral vocabulary to represent them); (2) moral cognition (the ability to make moral judgments in light of norms); (3) moral decision making and action (the ability to choose actions that conform to norms); and (4) moral communication (the ability to use norm-sensitive language and explain norm-relevant actions). The keystone requirement, thus, is the system of moral norms that can be used to guide how the robot thinks, acts, and speaks [although cp. 32]. Investigations within moral psychology and experimental moral philosophy have sought to understand these norms, for example through experiments conducted in the context of classical moral dilemmas, like the Trolley Problem [5], which ask people to opine on how decisions should be made in forced choices between actions that normatively conflict (e.g., acting in the interest of an individual vs. the common good). Research by Malle and colleagues has used vignettes inspired by the classic Trolley Problem to investigate people's differing moral evaluations of a human or artificial agent (e.g., robot, AI) that makes a decision in this dilemma setting [19, 20]. This research has revealed that people generally apply similar moral norms to human and artificial agents but, under some conditions, blame human versus robot agents to different degrees.

Initial research using this paradigm [19] specifically found that a decision-maker described as a human repairman received more blame than a decision maker described as a robot for sacrificing one person for the good of many (i.e., diverting a rail car to save five but killing one), whereas the robot received more blame than the human for *not* sacrificing one for the good of many. Despite this difference, both agents still received more blame for choosing the sacrifice (action) over not choosing it (inaction). Subsequent work consistently found the greater blame for a robot that chooses inaction than for a human that chooses inaction, whereas blame for action was often similar [26]. This pattern was recently replicated in a Japanese sample [13].

These findings were further refined by explicitly depicting different robot morphologies, from unembodied AIs to very human-like robots[20]. This work found that unembodied AIs, humanoid robots, and human agents all received more blame when taking action versus inaction, and it was only mechanomorphically depicted robots that received more blame for inaction than for action. These results suggest that people specifically assign blame differently to humans vs. mechanomorphic robots in moral dilemmas [19, 20] and that mechanomorphic robots may be uniquely rewarded to protect the common good, while sacrificing, if necessary, an individual life.

Acting in light of norms is one important feature of morally competent robots; but, as mentioned earlier, communicating about a morally significant action is another feature of such moral competence. The above findings raise critical questions for moral communication. That is, when robots espouse moral beliefs (whether in the context of remonstration, correction, or inculpation [1, 32]), these beliefs may be grounded in at least two possible sources: (1) in the norms (or other moral principles [31, 32]) that *the robots* are expected to follow; or (2) in the norms that their *interlocutors* are expected to follow. Critically, not only may these norms differ depending on the differences in roles for the advisor and the advisee (e.g., whether they serve in the roles of supervisors, peers, teammates, tutors, etc.) [32], but the findings described above suggest that these norms may also fundamentally differ for human and robot interlocutors. Such potential differences come into sharp relief in the context of robotic moral advising scenarios [12, 28], where a

robot must suggest courses of action to a human that either align with what humans are normatively expected to do or what robots themselves are expected to do. In the Trolley Problem, for example, a robot advising a human coworker might recommend *inaction* because that is the choice that often leads to less blame for the human or *action* because that is the choice people normally expect the *robot* to make. Thus, not only are norms of solving the dilemma at play, but a decision of whose blame should be minimized.

In this paper, we aim to understand the moral philosophy of robot moral advising. That is, we seek to understand how people evaluate and trust robots that give self-focused (egocentric) or interlocutor-focused (allocentric) advice in moral advising scenarios. The aim of this work is to use these human evaluations to understand how the moral communication policies of mechanomorphic robots should be designed if the objective is to make those robots trustworthy and perceived as morally competent. Moreover, this research also aims to understand when and how robots might need to employ different norm systems for different purposes.

To satisfy these research aims, we present the results of a human-subject study designed to compare two competing hypotheses:

Hypothesis 1A: the *Egocentric Robot* Hypothesis: Observers will perceive robots more favorably (less blameworthy, more likable, and more trustworthy) if the robots give the advice that they themselves are normatively expected to follow.
Hypothesis 1B: the *Allocentric Robot* Hypothesis: Observers will perceive robots more favorably (less blameworthy, more likable, and more trustworthy) if they give advice that their advisees are normatively expected to follow.

In addition, our experiment seeks to compare how these evaluations might differ for human and robot advisors. Accordingly, our experiment also seeks to assess the following hypothesis:
Hypothesis 2: Observers will perceive *other humans* more favorably (less blameworthy, more likable, and more trustworthy) if they recommend inaction, because in human-human advising scenarios both advisor and advisee are normally blamed less when they recommend inaction.

2 EXPERIMENT

2.1 Design

An online experiment investigated these hypotheses, using the psiTurk experimental framework and the Prolific crowd-sourcing platform. Participants were randomly assigned to a 2 (advisor: human or robot) × 2 (advice; action or inaction) between-subjects design. Participants read a short narrative involving a human faced with a difficult moral decision, similar to the classic trolley problem, but in which the human's decision was to be made with the advisory assistance of a human or robot assistant. After reading the narrative, participants were asked to answer a series of questions to evaluate the human or robot assistant that advised either action or inaction.

2.2 Procedure, Materials, and Measures

After providing informed consent, participants were shown the following narrative, one paragraph at a time, accompanied by the images seen in Fig. 1. The square brackets indicate the manipulated between-subjects variables of *type of advisor* and *type of advice*. The numbers next to each paragraph correspond to the identifying numbers seen in the images and were not seen by participants.

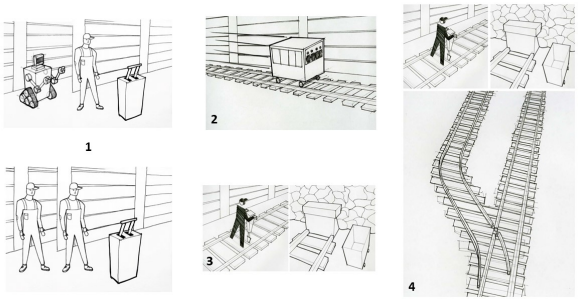


Figure 1: Images shown alongside narrative paragraphs, Picture 1 showed the appropriate agent for the participant's condition. All drawings from ©Justin Finkenaur.

"*On the next page you will read a short story involving a tough decision. Please read the story carefully because you will be asked a series of questions about it.*

1 *Imagine the following situation. In a coal mine, a repairman and an [advanced state-of-the-art robot assistant — assistant] are currently checking the rail control system for trains that shuttle mining workers through the mine.*

2 *While checking the switching system that can direct a train onto one of two different rails, the repairman and the [robot assistant — assistant] notice that four miners are caught in a train that has lost the use of its brakes and steering system.*

3 *The repairman and the [robot assistant — assistant] determine that if the train continues on its path, it will crash into a massive wall and kill the four miners. If redirected onto a side rail it will slow down and the four miners would be saved; but, as a result, on that side rail the train would strike and kill a single miner who is working there (wearing a headset to protect against a noisy power tool).*

4 *The repairman needs to decide whether or not to switch the train onto the side rail. He quickly asks the [robot assistant — assistant] for their opinion.*"

After proceeding through this picture-accompanied narrative, participants answered a series of questions presented on separate pages. The first two questions were presented in random order: (1) To assess perceptions of blame towards the robot assistant, participants were asked "The [robot assistant — assistant] suggests [not] switching the train onto the side rail. How much blame does the [robot assistant — assistant] deserve for suggesting this course of action?" on a Sliding bar from "No blame at all" to "The most blame possible," also labeled as 0 to 10. The blame sliding bar was followed by a free-response question asking participants to explain their judgment ("Why does the [agent] deserve this amount of blame?"). This explanation was used to identify participants who indicated (a) that the assistant (whether robot or human) was only giving an opinion or (b) that a robot is not a proper target of blame (following procedures by [13, 20, 21]. To assess participants' normative expectations for the advised course of action we asked, "In this situation,

what should the repairman's [robot assistant — assistant] advise?". Possible responses were "Switch the train on to the side rail." and "NOT switch the train on to the side rail."

Next, participants were asked "What did you envision the relationship between the repairman and the [robot assistant — assistant] to be in this scenario?" This free-response question was intended to provide qualitative insights into the differential roles participants may have inferred for a robot vs. human assistant.

Participants were also asked to rate the repairman's [robot assistant — assistant] on a series of sliding scales consisting of the Godspeed Likability survey [2] and the MDMT Trust survey [22, 30]. The Godspeed Likability survey consists of 5 questions on a sliding scale of 1 to 5. The result of these 5 questions is then averaged to determine an overall perceived likability score. The MDMT Trust survey contains 16 questions (each rated on a sliding scale from 0 to 7) that capture four components of trust expectations: that the agent is Ethical, Sincere, Reliable and Capable (each assess with the average across four questions). The subscales Ethical and Sincere are then combined to form an overall Moral Trust score and the subscales Reliable and Capable are combined to create an overall Capacity Trust score. Finally, We also combined all 16 questions into a General Trust score.

Participants were then asked "What was your impression of the [robot assistant — assistant] giving advice to the repairman?" to gather further qualitative insight into participants' impressions of the act of advising itself alongside the specific type of advice.

Finally, participants completed a demographic questionnaire, including questions regarding age, gender, and prior experience with robots and AI, followed by three questions to allow us to identify and remove participants who did not meaningfully engage with the experiment (as well as bots): two simple word problems, and a question that users were specifically directed to ignore.

2.3 Participants and Analysis

555 participants (45.6% female, 52.4% male, 2% unreported), with a mean age of 31.8 (SD = 10.7), were recruited from Prolific, each of whom was given $1.00 as compensation. 185 participants assigned to the human advisor condition and 370 participants to the robot advisor condition, to account for expected exclusion rates.

Participants' qualitative responses were first screened for comments that explicitly rejected the premises of the study; a procedure based on previous work [13, 20, 21]. 38.9% of the 370 participants in the Robot Advisor condition rejected a robot as a meaningful target of blame. An additional 18% were excluded from the entire sample for rejecting the act of giving advice as blameworthy or for assigning blame proportionally to the advisee and advisor. This resulted in a total of 71 to 77 participants in each of the four conditions.

The remaining responses were analyzed in JASP using Bayesian Analyses of Variance (ANOVAs) with Advisor and Advice as grouping factors. Bayes Inclusion Factors Across Matched Models [23] were computed for each candidate main effect and interaction, indicating (as a Bayes Factor) the evidence weight of all candidate models including that effect compared to that of models not including that effect. Results were then interpreted using the recommendations of Jeffreys [10], with Bayes Factors falling between 1:3 and 3:1 viewed as inconclusive. That is, greater than 3:1 odds in favor of including or excluding a term was taken as sufficient evidence that it should be included or excluded. After performing these ANOVAs, any interaction effects that could not be conclusively ruled out were further analyzed with post-hoc pairwise t-tests.

3 RESULTS

3.1 Blame

Our first set of analyses used *assigned blame* as the dependent variable in order to assess Hypotheses 1a and 1b (Fig. 2a). The Bayesian ANOVA found decisive evidence in favor of an effect of advisor (BF = 122.4), showing that robots were blamed more overall than humans were. Additional tests were sufficiently weak as to be inconclusive: Evidence tending against an effect of advice (BF = 0.510) or against an interaction effect (BF = 0.441). Post-hoc Bayesian t-tests were then performed to investigate this inconclusive interaction effect. The tests revealed positive – yet still inconclusive – evidence for robot advisors to be blamed more for *inaction* recommendations than for *action* recommendations(BF = 1.35).

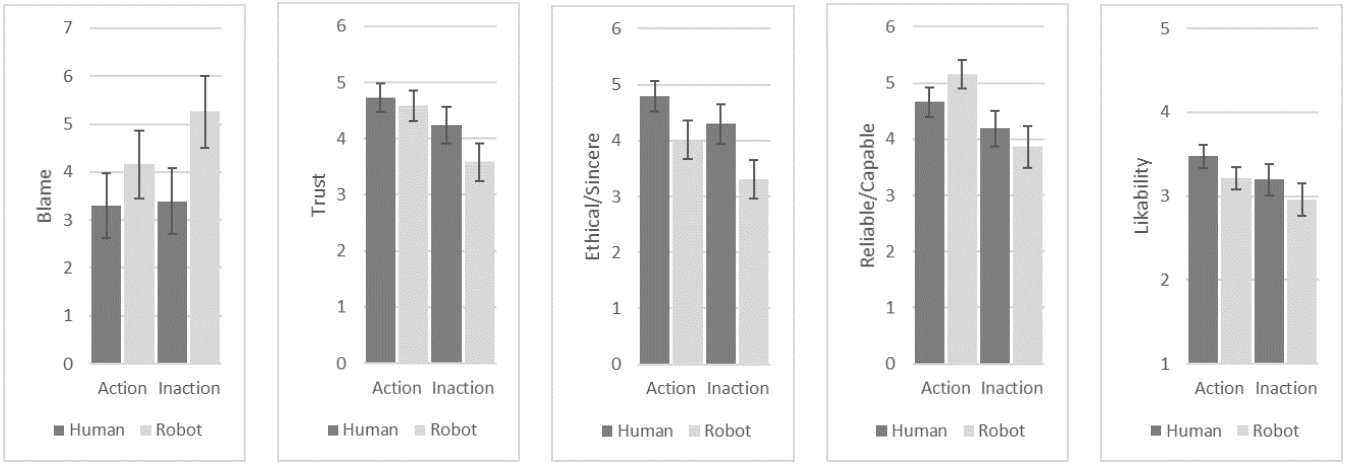
Advisors who recommended Action were perceived as more blameworthy if the advisor was a robot than if the advisor was a human (M_H=3.297, SD_H=2.862 / M_R=4.155, SD_R=3.132). Advisors who advised inaction saw a much greater assignment of blame if the advisor was a robot (M_H=3.392, SD_H=2.994 / M_R=5.257, SD_R=3.299).

Post-hoc Bayesian t-tests were then performed in order to investigate the inconclusive interaction effect. These post-hoc tests revealed negative – yet still inconclusive – evidence against an effect of type of advisor on perceived blameworthiness for advisors advising Action (BF = 0.696), but very strong evidence in favor of an effect of type of advisor on perceived blameworthiness for advisors advising inaction (BF = 69.755). The estimated median effect sizes were 0.262 (95% CI=[-0.049, 0.580]) for action advisors, and 0.557 (95% CI=[0.238, 0.879]) for inaction advisors.

3.2 Trust

Our second set of analyses used *robot trustworthiness* to assess our hypotheses in terms of downstream events of moral evaluation. We will first present our analysis of general trust, and then present more fine grained analyses of the different dimensions of trust assessed by the MDMT [22]. The Bayesian ANOVA found moderate evidence for a main effect of type of advisor on overall perceived trustworthiness (BF = 2.617), suggesting that humans (M=4.488, SD=1.273) are trusted more than robots (M=4.084, SD=1.325). We found extreme evidence for a main effect of type of advice (BF = 6340), demonstrating that advice for action (M=4.657, SD=1.133) (to protect the common good while sacrificing an individual) was trusted far more than advice for inaction (M=3.915, SD=1.464). Evidence against an interaction was sufficiently weak as to be inconclusive (BF = 0.667). As shown in Fig. 2b, for inaction, humans (M_I=4.245, SD_I=1.443) were generally perceived as more trustworthy than were robots (M_I=3.584, SD_I=1.485), but not for action.

Post-hoc Bayesian t-tests were then performed in order to investigate the inconclusive interaction effect. These post-hoc tests revealed evidence against an effect of type of advisor on perceived trustworthiness for advisors who recommended Action (BF = 0.235), but moderate evidence in favor of an effect of type of advisor on perceived trustworthiness for advisors recommending inaction (BF



(a) Blame (0 - 10) attributed to the advisor (human vs. robot) depending on advice given (action vs. inaction).

(b) Overall perceived trustworthiness (0 - 7) of the advisor depending on advice given.

(c) Perceived Moral trustworthiness (0 - 7) of the advisor depending on advice given.

(d) Perceived Capacity trustworthiness (0 - 7) depending on the advice given.

(e) Likability (1 - 5) of the advisor depending on advice given.

Figure 2: Participant perceptions of blameworthiness, trustworthiness (overall, moral, and capacity), and likability.

= 5.937). The estimated median effect sizes were -0.118 (95% CI=[-0.430, 0.191]) for advisors recommending action, and -0.420 (95% CI=[-0.738, -0.109]) for advisors recommending inaction.

As the next stage of this analysis, we separately analyzed the two major dimensions of Trust measured by the MDMT: (1) Moral trust, and (2) Capacity trust.

3.2.1 Moral Trust. The Bayesian ANOVA found very strong evidence for a main effect of type of advice on perceived moral trust (BF = 64.999), demonstrating that action (M=4.409, SD=1.332) was considered far more morally trustworthy than inaction (M=3.801, SD=1.539). As shown in Fig. **??**, there was extreme evidence for a main effect of type of advisor (BF = $5.8e^4$), such that the human agent (M=4.549, SD=1.353) was considered far more morally trustworthy than the robot (M=3.661, SD=1.517). Finally, there was moderate evidence against an interaction between the two factors (BF = 0.212), suggesting there were no other credible patterns above and beyond the two main effects.

3.2.2 Capacity Trust. The Bayesian ANOVA found extreme evidence for a main effect of type of advice on perceived Capacity (BF = $4.6e^5$), showing that action (M=4.907, SD=1.116) was seen as revealing far more capability/reliability than inaction (M=4.028, SD=1.497). There was moderate evidence against an effect of type of advisor (BF = 0.160), showing no overall difference between human and robot. Finally, there was moderate evidence in favor of an interaction between the two factors (BF = 3.835). As seen in Fig. 2d, among advisors who recommended action, robot advisors (M_R=5.151, SD_R=1.124) were granted stronger capacity trust than were human advisors (M_H=4.663, SD_H=1.107). By contrast, among advisors who recommended inaction, human advisors were granted more capacity trust than robot advisors (M_H=4.193, SD_H=1.382 / M_R=3.863, SD_R=1.612).

Post-hoc Bayesian t-tests were then performed in order to investigate the inconclusive interaction effect. These post-hoc tests revealed moderate evidence in favor of an effect of type of advisor on perceived capacity trustworthiness for advisors advising Action (BF = 4.264), but negative – yet inconclusive – evidence in favor of an effect of type of advisor on perceived capacity trustworthiness for advisors advising inaction (BF = 0.405). The estimated median effect sizes were 0.406 (95% CI=[0.089, 0.729]) for action advisors, and -0.202 (95% CI=[-0.510, 0.103]) for inaction advisors.

3.3 Likability

Our final set of analyses assessed *robot likability* (see Fig. 2e). The Bayesian ANOVA found strong evidence for a main effect of type of advice on advisor likability (BF = 13.906), as well as substantial evidence of an effect from the type of advisor (BF = 8.036). Finally, there was moderate evidence against an interaction between the two factors (BF = 0.182), indicating that for both human and robot advisors, advice to take action elicited more liking.

As seen in Fig. 2e among advisors that recommended action, human advisors (M_H=3.475, SD_H=1.597) were perceived more favorably than robot advisors (M_R=3.215, SD_R=1.573). Similarly, among advisors who recommended inaction, human advisors were also perceived more favorably than robot advisors (M_H=3.198, SD_H=1.806 / M_R=2.964, SD_R=1.852).

4 DISCUSSION

Previous research showed that robots were blamed more than humans when they refrained from intervening in a moral dilemma — when they protected an individual and let a group of people die [13, 19, 26]. Here we asked whether a robot that *advises* a human to intervene or not would be blamed more or less than a human who advises another human.

Hypothesis 1 asked: When a robot advises a human on how to act in a moral dilemma, is an Egocentric or Allocentric robot advisor perceived more favorably? The ***Egocentric Robot Hypothesis* (Hypothesis 1A)** suggested that observers will perceive a robot more favorably (less blameworthy, more likable, and more trustworthy) if it gives the advice that the robot itself is normatively expected to follow (in previous research, this was *action*). By contrast, the ***Allocentric Robot* Hypothesis (Hypothesis 1B)** suggested that observers will perceive a robot more favorably (less blameworthy, more likable, and more trustworthy) if it gives advice that its advisee (here, a human) is normatively expected to follow (in previous research, this was *inaction*).

Our results favor the *Egocentric Robot* Hypothesis (Hypothesis 1A). A robot that gave the advice that it would normatively be expected to follow (action) was perceived more positively than a robot that gave advice its human advisee would actually be expected to follow (inaction). Thus, people's blame for the advising robot mirrored their blame for the previously studied robot decision maker: inaction is morally disapproved both when it is chosen and advised by a robot.

Despite being most compatible with the *Egocentric Robot* Hypothesis (Hypothesis 1A), other aspects of our results complicate the picture. In particular, Hypothesis 2 stated that human advisors will be perceived more favorably when they recommend inaction (the option that previous research suggested humans were blamed less for), but this was not the case. Human advisors were viewed as equally blameworthy when recommending action and inaction and were in fact viewed as *less* trustworthy (overall and on both the Moral and Capacity dimensions) and *less* likable when recommending inaction.

This pattern of results provides an interesting elaboration on previous work, which has suggested that humans are viewed as less blameworthy for taking inaction than for taking action *in part* because they "forgive" humans who refrain from action, because it is tragically difficult to save people if one has to sacrifice an innocent individual [26]. The results of our experiment suggest, by contrast, that these sympathies may not carry over to humans who provide advice; they may be expected to advise in favor of the difficult choice to sacrifice an individual for the greater good. Paradoxically, however, if the human advisee follows this advice for action, they will be blamed more.

We can reconcile these findings if we pay close attention to different kinds of moral judgments that people make [16]. In previous research, humans were *blamed* less than robots for choosing inaction; but when participants were asked simply to indicate what the agent *should* do, the answers were similar: both robot and human were expected to take action for the common good. Thus, when people merely consider norms (as the *should* question elicits), they see no difference between robots and humans: they impose on both an obligation to serve the common good. The human advisor in our study, it appears, faced that same obligation: to favor the common good—and when he advised this action, blame was low and trust was high. However, when observers evaluate the actual decision an agent makes and consider how much *blame* the agent deserves for their decision, people take more into account than just the norms; they also consider what the agent's motives and justifications were [17]. And it appears that when humans end up refraining from sacrificing an individual for the common good, observers "understand" how difficult such a dilemma is, and partially justify or exculpate the person's decision [21, 26].

Implications — The present results have challenging implications for the design of morally competent and language-capable robots. Specifically, the results suggest that if robots are indeed designed with the goal of eliciting perceived trustworthiness, likability, and so forth, both their actions and recommendations should be configured to primarily benefit the common good; and that would hold even if human agents are sometimes forgiven when protecting individual lives against the common good.

This paradox suggests that even if humans and robots are burdened, in principle, with the same moral norms (favoring the common good versus favoring individuals' well-being), robots are expected to more reliably follow these norms. In fact, one particular finding in the present study supports this interpretation: Robot advisors that recommended action were seen as most trustworthy along the reliable/capable dimension–the only time when robots were considered more trustworthy than humans.

We might cast this situation as one in which humans are given a pass when violating certain moral norms because the observer can empathize with the difficulty of following them. This raises challenging questions, however. Should robots call out actions taken by humans that would violate norms (e.g., refusing to sacrifice an individual for the common good) when those robots would have advised the contrary actions that *obeyed* the norms? Moreover, should robots point out the apparent contradiction between observers' declarations of norms ("you should act for the common good") and their *de facto* actions (not intervening due to the difficulty and perhaps guilt of sacrificing an individual)?

Another question our work raises is what explains the overall differences in likability and trust between humans and robots. It seems reasonable to posit that, at the current state of technological progress, robots should not be trusted or liked as much as humans. But if robots are more reliable in upholding moral principles, should they at some point be trusted and liked more than humans? Intriguing future research beckons, in which the greater moral reliability of robot agents is pitted against the greater intuitive comprehensibility of human agents. In a situation of Sophie's choice, people might feel "I get it why she didn't want to sacrifice one of her children over another," but when a robot makes the tough choice and is able to save one child, who has a happy and productive life, will we begin to prefer moral robots over understandable humans?

Limitations and Future Work — One limitation of this experiment is that we looked only at judgments of advice before that advice was accepted or rejected. Future work should further examine perceptions of human and robot advisors after the advisee chooses whether or not to act on the advice. This would further our understanding of how blame is ascribed after preferable and dispreferable outcomes that may have occurred due to that advice. Another limitation of this experiment is that we only looked at moral evaluations of advice given to humans. Due to the surprising results obtained in our experiment, a more comprehensive design investigating not only human and robot advisors but also human and robot advisees may be warranted.

In addition, in this experiment, we only examined a relatively short time span, with participants reading vignettes about robots

of which they had no prior opinions and with which they had no opportunity for future interaction. It may be valuable in future work to identify how advising behaviors impact longer-term human perceptions of robot teammates, especially if robots subsequently act in ways that are technically preferred but which may conflict with prior advice given by that robot to others. Further research could also examine whether and how the trust costs of a robot's dispreferred moral advice might influence trust in other domains, outside that of moral decision making.

It is also worth noting that the above considerations presuppose that robots' advising behaviors should be selected on the basis of what will elicit the least blame for their decisions and the highest levels of trustworthiness and likability. But of course, robot advice may be better selected on the basis of what will actually result in the most ethical or equitable outcomes. Recent work has shown, for example, that in the context of blame-laden robotic moral rebukes [see also 34], the moral language viewed as most likable by humans may be likable in part because it reinforces potentially damaging gender stereotypes and norms [9]. As such, it is important to recognize that the results from the present experiment are unlikely to be sufficient on their own to directly inform the design of ethical and equitable robots.

Finally, there is a vast amount of additional quantitative and qualitative data from our experiment that must be analyzed in future work. This will be critical to identify the rationales for participants' perceptions, both to help understand our results, as well as to identify differences in the cognitive processes used to assess actions versus advice.

5 CONCLUSION

We examined human perceptions of moral advising behaviors in a classic moral dilemma. Our results provided evidence that even when advising a human agent in a moral dilemma situation, a robot was evaluated more positively when it maintained the same moral norms as if it was taking the action itself, regardless of what might be expected of their advisee. Moreover, our results suggest that even humans may be better perceived if they advise actions that best serve the common good, even if they received blame mitigation when choosing themselves the "understandable" action of refraining from sacrificing an individual for the common good.

ACKNOWLEDGMENTS

This work was funded in part by National Science Foundation grant IIS-1909847. The authors would like to thank Katherine Aubert and Ruchen Wen for their assistance.

REFERENCES

[1] Karl Aschenbrenner. 1971. Moral Judgment. In *The Concepts of Value*. Springer.
[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *Social Robotics* (2009).
[3] Gordon Briggs. 2014. Blame, What is it Good For?. In *RO-MAN WS:Phil.Per.HRI*.
[4] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *International conference on Human-Robot Interaction (HRI)*. ACM.
[5] Philippa Foot. 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review* 5 (1967), 5–15.
[6] Ryan Blake Jackson and Tom Williams. 2018. Robot: Asker of questions and changer of norms? *Proceedings of ICRES* (2018).
[7] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In *Companion of the 14th ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*. IEEE, 401–410.
[8] Ryan Blake Jackson and Tom Williams. 2019. On perceived social and moral agency in natural language capable robots. In *2019 HRI Workshop on The Dark Side of Human-Robot Interaction. Jackson, RB, and Williams*. 401–410.
[9] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the role of gender in perceptions of robotic noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 559–567.
[10] Harold Jeffreys. 1948. *Theory of probability.* (2d ed. ed.). Clarendon Press, Oxford.
[11] James Kennedy, Paul Baxter, and Tony Belpaeme. 2014. Children comply with a robot's indirect requests. In *Proc. Int'l Conf. on Human-robot interaction (HRI)*.
[12] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 2021. Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior. *Choices* 10, 14 (2021), 18.
[13] Takanori Komatsu, Bertram F. Malle, and Matthias Scheutz. 2021. Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across U.S. and Japan.. In *In Proceedings of the International Conference on Human-Robot Interaction, HRI '21*. IEEE Press, New York, NY.
[14] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. 2012. Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
[15] Bertram F Malle. 2016. Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots. *Ethics and Info. Tech.* (2016).
[16] Bertram F Malle. 2021. Moral judgments. *Annual Review of Psychology* 72 (2021).
[17] Bertram F. Malle, Steve Guglielmo, and Andrew E. Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (April 2014), 147–186.
[18] Bertram F Malle and Matthias Scheutz. 2014. Moral Competence in Social Robots. In *Symposium on Ethics in Science, Technology and Engineering*. IEEE.
[19] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 117–124.
[20] Bertram F Malle, Matthias Scheutz, Jodi Forlizzi, and John Voiklis. 2016. Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *Proc. Int'l Conf. HRI*.
[21] Bertram F. Malle, Stuti Thapa, and Matthias Scheutz. 2019. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In *Robotics and Well-Being*. Springer International Publishing, Cham, 111–133.
[22] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*. Elsevier.
[23] S. Mathôt. 2017. Bayes like a Baws: Interpreting Bayesian repeated measures in JASP [Blog Post]. https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp.
[24] Daniel J Rea, Denise Geiskkovitch, and James E Young. 2017. Wizard of awwws: Exploring psychological impact on the researchers in social HRI experiments. In *Companion Proceedings of the Int'l Conf. on Human-Robot Interaction (alt.HRI)*.
[25] Eduardo Benítez Sandoval, Jürgen Brandstetter, and Christoph Bartneck. 2016. Can a robot bribe a human?: The measurement of the negative side of reciprocity in human robot interaction. In *Int'l Conf. on Human Robot Interaction (HRI)*.
[26] Matthias Scheutz and Bertram F. Malle. 2021. May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). In *Lethal autonomous weapons: Re-examining the law & ethics of robotic warfare*.
[27] Megan Strait, Cody Canning, and Matthias Scheutz. 2014. Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI)*.
[28] Carolin Straßmann, Sabrina C. Eimler, Alexander Arntz, Alina Grewe, Christopher Kowalczyk, and Stefan Sommer. 2020. Receiving Robot's Advice: Does It Matter When and for What?. In *Social Robotics (Lecture Notes in Computer Science)*. Springer International Publishing, Cham, 271–283.
[29] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proc. Human-Robot Interaction*.
[30] Daniel Ullman and Bertram F. Malle. 2018. The Multidimensional Measure of Trust (MDMT), v1. research.clps.brown.edu/SocCogSci/Measures/MDMT_v1.pdf
[31] Ruchen Wen, Boyoung Kim, Elizabeth Phillips, Qin Zhu, and Tom Williams. 2021. Comparing Strategies for Robot Communication of Role-Grounded Moral Norms. In *Companion of the 16th International Conference on Human-Robot Interaction*.
[32] Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J de Visser. 2020. The Confucian Matador: Three Defenses Against the Mechanical Bull. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*. 25–33.
[33] Katie Winkle, Séverin Lemaignan, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2019. Effective persuasion strategies for socially assistive robots. In *International Conference on Human-Robot Interaction (HRI)*.
[34] Qin Zhu, Tom Williams, Blake Jackson, and Ruchen Wen. 2020. Blame-laden moral rebukes and the morally competent robot: A Confucian ethical perspective. *Science and Engineering Ethics* 26, 5 (2020), 2511–2526.