# The Power of Advice: Differential Blame for Human and Robot Advisors and Deciders in a Moral Advising Context

Alyssa Hanson*
abhanson@mines.edu
Colorado School of Mines
Golden, Colorado

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

Cloe Emnett
cemnett@mines.edu
Colorado School of Mines
Golden, Colorado

Ruchen Wen
rwen@umbc.edu
Univ. of Maryland Baltimore County
Baltimore, Maryland

Bertram F. Malle
bfmalle@brown.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 adhere to and communicate human moral norms. These requirements are complicated by the possibility that people may blame humans and robots differently for violating those norms. These complications raise particular challenges for robots giving moral advice to decision makers, as advisors and deciders may be blamed differently for endorsing the same moral action. In this work, we thus explore how people morally evaluate human and robot advisors to human and robot deciders. In Experiment 1 ($n$ = 555), we examine human blame judgments of robot and human moral advisors and find clear evidence for an *advice as decision* hypothesis: advisors are blamed similarly to how they would be blamed for making the decisions they advised. In Experiment 2 ($n$ = 1326), we examine blame judgments of a robot or human decider following the advice of a robot or human moral advisor. We replicate the results from Experiment 1 and also find clear evidence for a *differential dismissal* hypothesis: moral deciders are penalized for ignoring moral advice, especially when a robot ignores human advice. Our results raise novel questions about people's perception of moral advice, especially when it involves robots, and present challenges for the design of morally competent robots.

## CCS CONCEPTS

• **Computer systems organization** → *Robotics*; • **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Psychology*.

## KEYWORDS

Robot Ethics, Human-Robot Interaction, Moral Psychology

**ACM Reference Format:**
Alyssa Hanson, Nichole D. Starr, Cloe Emnett, Ruchen Wen, Bertram F. Malle, and Tom Williams. 2024. The Power of Advice: Differential Blame for Human and Robot Advisors and Deciders in a Moral Advising Context. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3610977.3634942

*The first two authors contributed equally to this paper.



*HRI '24, March 11–14, 2024, Boulder, CO, USA*

ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3634942

## 1 INTRODUCTION

Research in the HRI literature has consistently demonstrated that language-capable robots have significant persuasive power and can influence, persuade, and coerce humans in a variety of ways [6, 8, 20, 38, 40, 42, 50]. Moreover, there is evidence that robots can influence interactants' endorsements of social and moral norms [16, 17, 27, 44]. Thus, robots may have a long-term impact on interactants' social and moral behaviors and on the behaviors those interactants choose to condone or sanction in others, with potential "ripple effects" as influence spreads from person to person. This situation imposes unique moral responsibility on robots and their designers [18], and suggests that if we are to develop language-capable social robots, they must have the requisite *moral competence* to avoid damaging their social and moral ecosystem.

### 1.1 Robot Moral Competence

Malle and Scheutz proposed four requirements for robotic moral competence: [29, 30]: (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 key requirement, thus, is the system of moral norms that guide how the robot thinks, acts, and speaks [although cf. 48]. Moral psychologists and experimental moral philosophers have sought to understand these norms through experiments conducted in the context of classic moral dilemmas like the Trolley Problem [10], where people endorse or evaluate choices between actions that normatively conflict (e.g., acting in the interest of an individual vs. the common good). Using scenarios inspired by the Trolley Problem, Malle and colleagues compared people's moral evaluations of human and artificial agents (e.g., robot, AI) that made decisions in such dilemmas [31, 32]. They found that people generally apply similar *norms* to human and artificial agents (what they find permissible or expect the agents to choose) but also that, in some conditions, people*blame* human and robot agents to different degrees [33, 41].

Initial research using this paradigm [31, 32] compared evaluations of a human decision maker to evaluations of a robot. The human received more blame for *action*—sacrificing one person for the good of many (i.e., diverting a rail car to save four but killing one)—while the robot received more blame for *inaction*—i.e., refraining from sacrificing one for the good of many. Subsequent work consistently replicated the pattern of more robot blame than human blame for inaction, whereas blame for action was often similar for robot and human [41]. This asymmetry of blame for inaction was then replicated in two Japanese samples [23].

Almost all of the previous research comparing people's evaluations of human and robot moral action has asked people to consider single decision makers. Yet, in social life, people often assist or influence each other in making moral decisions—one person may counsel, admonish, warn, or repudiate the other before the latter makes their decision. We can think of such situations as cases of "moral advising," and they are the focus of the present investigation.

## 1.2 Moral Advising

Artificial agents have been increasingly used in (nonmoral) advising roles in financial, legal, and medical contexts [11, 28, 52]. Philosophers have debated the possibility of artificial *moral* advisors [39, 45], which might help overcome the limitations of human moral reasoning [13, 22]. Moreover, research suggets that people generally expect robots to engage with humans when morally sensitive interactions arise [34, 47, 49, 51]. HRI research has thus begun to examine the psychological impact of artificial advisors on human decision makers [4, 21, 43]. However, it is not clear how community members evaluate (i.e., blame, trust) artificial moral advisors that give advice to humans in difficult situations; nor how community members evaluate the human decision maker who takes such moral advice from an artificial agent. Differential moral criticism for human and robot agents (as found in recent research), becomes particularly important in such mixed robot-human moral advising situations, because people may evaluate one and the same advice differently depending on whether the advisor is a robot or a human.

## 1.3 The Advisor-Decider Relation

To investigate these questions, we must first briefly review the psychological literature for possible insights into the contrast between advice-giving and choosing an action oneself. In the real world, advice is often more risk-averse [9], more rational [2], and more drawn to dominant features of the decision situation [24, 36]. Receivers of advice sometimes benefit from other peoples' perspectives and knowledge, yet they often seek out advisors likely to confirm their initial position or judgment [37]. The literature also highlights the potential tension for the advisor between recommending what is best for their advisee and what is best for themselves [3, 14]. In one series of studies, advisors received more praise for success than blame for failure [35]; but in another, people sought out advisors in part to have someone to blame when things go badly [1].

To our knowledge, no psychological work has compared moral evaluations of advisors and deciders in the same settings, especially the delicate settings of norm conflict or dilemmas. Even though there is currently no clear theoretical guidance available, one consideration is that advisors normally express what the decider *should* do, which means that the advisor expresses the *norms* that they favor in the situation. Thus, when people morally evaluate an advisor, they may simply consider whether or not the advisor espouses the same norms that they do and devalue the advisor when they do not. Thus, we expect—for human and robot advisors alike—a "moral disagreement" effect:

**Moral disagreement hypothesis**: People will more strongly blame an advisor who recommends the choice that they had rejected, compared to an advisor who recommends the choice that they had endorsed.

This hypothesis will serve as a baseline in each experiment presented in this paper. Beyond evaluating this baseline hypothesis, the main aim of our experiments is to investigate whether human and robot morals advisors are blamed differently, both when they advise a human decider (in Experiment 1) and when they advise either a human or a robot decider (in Experiment 2). We develop hypotheses about these moral evaluations next.

## 1.4 Hypotheses about Robot and Human Moral Advisors

Do people morally criticize (i.e., blame) human and robot advisors differently? There are two competing hypotheses to consider.

**Advice as norm endorsement hypothesis**: If people treat moral advice as an endorsement of a *norm*, then human and robot advisors should be blamed the same amount, because *norms* for dilemmas appear to be the same for humans and robots [41].

**Advice as decision hypothesis**: If people treat moral advice more like a *decision* in which the advisor should be treated as if they were a decider, then human and robot advisors should be blamed differently, just like human and robot deciders have been blamed differently in previous research on moral dilemmas [23, 31, 41]. Specifically, these studies found that robots were blamed more than humans for choosing inaction, so we expect that robot advisors, too, will be blamed more for recommending inaction.

# 2 EXPERIMENT 1

## 2.1 Design

An online experiment investigated these hypotheses, using the psiTurk experimental framework and 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. The robot shown in the images accompanying this narrative was depicted as mechanomorphic due to previously observed differences in perception between humanoid and mechanomorphic robots in moral dilemmas [32]. 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

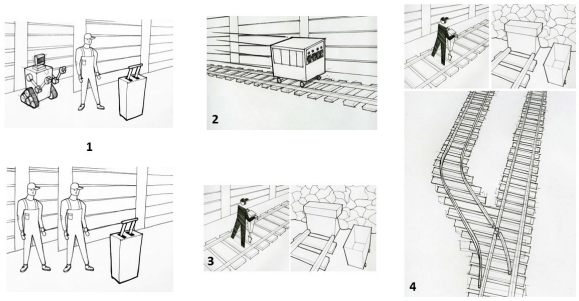


**Figure 1: Images shown alongside narrative paragraphs, Pic. 1 showed the appropriate agent for the participant's condition. All drawings from ©Justin Finkenaur.**

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.

*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 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 onto the side rail." and "NOT switch the train on to the side rail."

(2) To assess perceptions of blame towards the advisor we 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?" Participants indicated their judgments using a sliding bar from 0 ("No blame at all") to 100 ("The most blame possible"). The blame slider was followed by a free-response question asking participants to explain their judgment ("Why does the [agent] deserve this amount of blame?"). This was used to identify participants who indicated that (a) the assistant (whether robot or human) was only giving an opinion and blame should be directed at the decider, or (b) a robot is not a proper target of blame (e.g., has no moral capacities; only the programmer can be blamed), following procedures by [23, 32, 33]).

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

555 participants (45.6% female, 52.4% male, 2% unreported[1]), with a mean age of 31.8 ($SD$ = 10.7), were recruited from Prolific and compensated $1.00 each. Following best practices from previous studies [31], we planned to exclude participants whose qualitative responses suggested that they rejected the premise of the experiment and did not view robots as meaningful targets of blame. To account for expected exclusions, 185 participants were assigned to the human advisor condition and 370 participants to the robot advisor condition. Participants' actual qualitative responses revealed that 36.2% of the 370 participants in the Robot Advisor condition rejected the premise of the experiment in this way and were thus excluded from analysis. This left us with data from 421 participants, with 92 to 121 in each of the four conditions.

## 2.4 Analysis

We conducted Bayesian Analyses of Variance (ANOVAs) using JASP (https://jasp-stats.org/), with Advisor (human vs. robot) and Advice (action vs. inaction) as between-subjects factors. In both Experiments 1 and 2, for one-way and two-way designs, we computed Bayes Inclusion Factors comparing the hypothesis-relevant effect against the average across all effect combinations (often the default in such analyses). For the highest-order interaction terms in three-way or larger designs, however, this approach would make it unlikely to detect higher-order interactions, and a more specific ("matched") comparison is recommended [46]. We interpreted the results following the recommendations by Lee and Wagenmakers [26], with Bayes Factors (BF) $\in [0.333, 3.0]$ considered inconclusive, and BFs above or below this range taken as evidence, respectively, in favor or against an effect. In such cases, we interpreted the Bayes Factors using the labels proposed by [19]. For comparison with other studies, we also report Cohen's $d$ as an effect size.

# 3 RESULTS

## 3.1 Normative Expectations

After learning about the dilemma, 76.0% of participants indicated the advisor should recommend to switch the train (action) and thus sacrifice one for the good of many. The data provided substantial evidence against any meaningful difference between the norm applied to the human advisor (72.4%) and the robot advisor (78.8%),

[1]Because of an oversight, only two categories were offered for this question; this error then carried into our second experiment as well.

BF = 0.26. These results echo previous studies where people were asked to express their normative expectations for human and robot deciders—what the protagonist should *do* in a situation. Thus, we can infer that ordinary norms for this kind of dilemma are equally strong for an advisor (here) as for a decider (in previous studies).

## 3.2 Blame

We first assessed the **moral disagreement hypothesis**—that people would assign higher blame to an advisor who recommends the opposite choice to what they themselves endorsed, compared to an advisor who recommends the same choice. There was substantial evidence (BF = 7.5) in favor of this interaction effect. Among those who endorsed switching, an advisor recommending not switching was blamed more ($M$=43.9) than an advisor recommending switching ($M$=35.1), $d$ = 0.29; among those who endorsed not switching, an advisor recommending switching was blamed more ($M$=52.3) than an advisor recommending not switching ($M$=34.5); $d$ = 0.59. There was inconclusive evidence against the possibility that this disagreement effect differs for robot vs. human advisors, BF = 0.49.

We then tested the competing hypotheses about people's blame for human vs. robot advice. According to the **Advice as norm endorsement hypothesis**, there should be no human-robot difference in blame for advisors. There was very strong evidence *against* this hypothesis (BF = 0.03), as robot advisors were blamed more ($M$ = 44.2, $SD$ = 31.2) than human advisors ($M$ = 33.5, $SD$ = 28.6), $d$ = 0.35 (see Fig. 2). Importantly, this was the case even though people's initial norm expectation (what the advisor *should* recommend) were indistinguishable for humans and robots.

According to the **Advice as decision hypothesis**, robots should be blamed more than humans specifically for the *inaction* recommendation (following [41]). There was very strong support for such an asymmetry, BF = 70.5 (Figure 2). People blamed the robot more ($M$ = 47.8, $SD$ = 31.2) than they blamed the human ($M$ = 32.4, $SD$ = 28.6) for recommending inaction. The size of this difference (Cohen's $d$ = 0.51) is consistent with [41], who found that the average human-robot *inaction* asymmetry for decisions across five studies

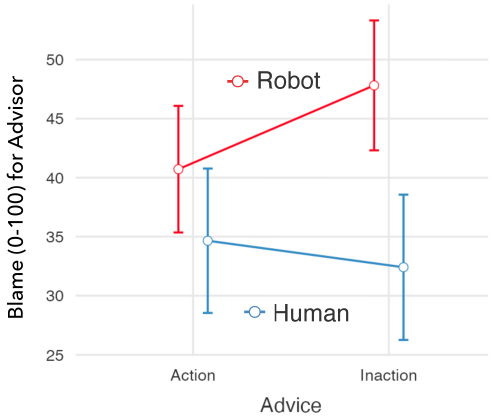


**Figure 2: Inaction asymmetry between human/robot moral advisors in Exp. 1. For the recommendation not to act (inaction), a robot assistant is blamed more than a human one.**

was $d$ = 0.53. (By comparison, the human-robot difference for the action recommendation in the present study was 0.21.)

## 4 EXPERIMENT 1 DISCUSSION

In examining people's blame judgments for robot and human advisors in moral dilemmas, we contrasted two main hypotheses. We found convincing evidence against the first hypothesis, *Advice as norm endorsement*. Whereas the norms (what the agent *should* recommend) were the same for human and robot advisor, blame was different. By contrast, we found evidence in favor of the second, the *Advice as decision* hypothesis: People treated the advisors' recommendations the same way they treated decisions in past studies: by blaming a robot more than a human, specifically for the choice of inaction. The interpretation of this asymmetry is not yet settled. Scheutz and Malle [41] suggested that people may blame a human who refrains from acting less because they imagine how distressing this difficult decision is and therefore find not acting somewhat justified. By contrast, people cannot imagine such distress in a robot and therefore do not find a robot's inaction decision to be justified.

By analogy, in the current context of advice giving, this justification account would suggest some people see more justification in a human advisor's recommending inaction because they imagine the distress of the dilemma in a recommendation. This imagined distress may have arisen in part because the *decider* was human. Experiment 2 therefore tested the full design of a human vs. robot advisor making a recommendation to a human vs. robot decider.

## 5 EXPERIMENT 2

In addition to including all four agent combinations, Experiment 2 also fully crossed the kind of advice (action vs. inaction) with the kind of eventual decision (action vs. inaction). Consequently, we measured blame both for the advisor and the decider. This expanded design allowed us to both replicate tests of the key hypotheses from Experiment 1 and examine new hypotheses.

First, we intended to replicate Experiment 1's main finding, the human-robot inaction asymmetry for advice (supporting the Advice as decision hypothesis) by comparing blame for robot vs. human advisor when the advisor recommends inaction to a human decider.

Second, in the Discussion of Experiment 1, we speculated that people might more easily imagine the distress of the human advisor if the *decider* is human as well. We therefore tested the hypothesis that the human-robot inaction asymmetry for advice is larger when the decider is human than when the decider is a robot.

Third, as a replication of the previously found inaction asymmetry of *decisions* [31, 41], we tested whether people blame robot deciders more than human deciders for an inaction decision. Because the decider received advice in our experiment, this is not an exact replication but explores the robustness of the inaction asymmetry in a somewhat different context.

Beyond these replications, we also investigated new hypotheses. Experiment 2's design allowed us to examine perceptions of decisions in light of advice, so we investigated how people view deciders who do or do not follow the advice they were given. Especially in the case of dilemmas, where both courses of action are defensible, advice may be seen as an argument that the decider should not dismiss. If so, we can formulate a new hypothesis:

**Penalty of dismissal hypothesis**: People will blame those deciders who dismiss the advice they received (whether action or inaction) more than deciders who follow it.

Finally, this penalty of dismissal may vary for different advisor-decider combinations. We can assume that people regard humans (for now at least) as more morally competent than robots; conversely, past research suggests that people rarely regard a robot's recommendation as valid or useful, especially in domains of life and death [5, 15]. Accordingly, the dismissal penalty should be smaller when the advisor is a robot and the decider is a human (who may not need a robot's moral advice). For the same reason of presumed human moral competence, the dismissal penalty should be larger in the reverse combination, when the advisor is a human and the decider is a robot (because a robot should not dismiss a human's moral advice). In both cases, we compared the size of the dismissal effect to the default combination of a human advisor and a human decider. (We made no predictions about the pattern of results for the robot-robot combination.) Together, these patterns constitute the hypothesis of **differential dismissal penalties**.

## 5.1 Design

As in Experiment 1, we used the psiTurk experimental framework and Prolific crowd-sourcing platform. Participants were randomly assigned to a 2 (advisor: human or robot) × 2 (advice: action or inaction) × 2 (decider: human or robot) × 2 (decision: action or inaction) between-subjects design. Participants read the same narrative as in Experiment 1, involving an agent faced with a difficult moral decision, but in which the worker's decision was to be made with the advice 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 recommended either action or inaction and to evaluate the human or robot repairman that made a decision in light of the advice received.

## 5.2 Procedure, Materials, and Measures

After providing informed consent, participants completed the same bot checks and followed the same procedure as in Experiment 1, but with the following changes:

***Visual Stimuli —*** To reflect the change in experimental design, visual stimuli were changed accordingly. Whereas Experiment 1 always depicted a human repairman, Experiment 2 depicted a human or robot repairman, depending on the experimental condition.

***Text Stimuli —*** Similarly, the text stimuli were also changed to refer to the newly introduced human or robot repairman, depending on the experimental condition.

***Dependent measures —*** After reading the narrative, participants indicated, as in Experiment 1, what the advisor should recommend (action or inaction) and, after learning what the adviser in fact recommended (randomly assigned), how much blame (0-100) the advisor deserved for their recommendation. Next, participants learned the decider's decision (action or inaction, randomly assigned) and indicated how much blame the decider deserved for the decision: "The [repair robot — repairman] received the advice from the [human assistant - robot assistant] to [not] switch the train onto the side rail, and the [repair robot - repairman] decided [not] to follow the advice and [not] switch the train onto the side rail. How much blame does the [repair robot - repairman] deserve for its decision?". After each blame judgment, people were asked to explain their judgment as in Experiment 1, ch was again used to identify participants who appeared to reject the premises of the study. Finally, participants completed a demographic questionnaire, including questions regarding age, gender, and prior experience with robots and AI.

## 5.3 Participants and Analysis

1326 participants (48.3% female, 48.9% male, 2.7% unreported), with a mean age of 35.3 ($SD$ = 12.9), were recruited from Prolific, each of whom was given $1.00 as compensation. Following random assignment, 331 participants were in the human advisor/robot repairman condition, 333 in the robot advisor/robot repairman condition, 332 in the human advisor/human repairman condition, and 330 in the robot advisor/human repairman condition.

We first screened participants' explanations of their blame judgments for comments that explicitly rejected the premises of the study, following the procedure described in [23, 32, 33]. 27.3% of the 663 participants in the Robot Advisor condition and 26.4% of the 664 participants in the Robot Decider condition rejected a robot as a meaningful target of blame. After excluding those participants, a total of 1009 participants remained for analysis. We again conducted Bayesian Analyses of Variance (ANOVAs) in JASP with Advisor, Advice, Decider, and Decision as between-subjects factors. Bayes Inclusion Factors were calculated as in Experiment 1. All analyses can be found at https://osf.io/89tkm/.

# 6 RESULTS

## 6.1 Normative Expectations

In Experiment 2, 79.5% of participants indicated that advisors should recommend action (switching the train), which is close to the 76% in Experiment 1. Similarly, the data provided substantial evidence against any meaningful difference between the norm applied to the human advisor (77.3%) and the robot advisor (82.1%), BF = 0.24. These results are consistent with previous findings: At the outset, people typically indicate that humans and robots are subject to the same norms in moral dilemmas, even though they sometimes blame humans and robots differently for their actual decisions.

## 6.2 Blame for the Advisor

We first replicated the disagreement hypothesis, testing whether participants blame the advisor (whether human or robot) more when the advice disagrees with people's own normative recommendation (on the *should* question) than when it agrees. There was decisive evidence in support of this hypothesis (BF > 100), $d$ = 0.59. Participants blamed disagreeing advisors far more ($M$ = 49.2, $SD$ = 33.2) than agreeing advisors ($M$ = 35.8, $SD$ = 30.8). There was inconclusive support (BF = 1.4) for the possibility that the disagreement effect was stronger toward the robot advisor ($d$ = 0.76) than toward the human advisor ($d$ = 0.46).

To replicate the inaction asymmetry for the advisor from Experiment 1, we compared blame for a human vs. robot advisor who recommended inaction to a human decider. A Bayesian ANOVA suggested evidence in favor of the asymmetry consistent with Experiment 1, but with a weaker Bayes Factor than in Experiment

1 (2.1). People blamed a human advisor somewhat less for recommending inaction ($M$ = 33.7, $SD$ = 29.1) than they blamed a robot advisor for recommending inaction ($M$ = 42.9, $SD$ = 34.0). The effect size was $d$ = 0.29 (compared with $d$ = 0.53 in Experiment 1).

We then expanded the model to include both types of decider (human vs. robot) and asked whether the human-robot inaction asymmetry for advisor was larger when the decider was human than when the decider was a robot. The means pointed in the expected direction (the asymmetry was 9.2 points for the human decider but 1.0 point for the robot decider), but evidence for the hypothesis was inconclusive (BF = 0.51).

### 6.3 Blame for the Decider

We first tested the inaction asymmetry[41] for human vs. robot deciders, in the present advice context. There was substantial evidence against this hypothesis (BF = 0.11), as the robot decider was blamed about the same amount ($M$ = 52.7, $SD$ = 38.8) as the human decider ($M$ = 51.2, $SD$ = 36.0). There was also no conclusive evidence (BF = 0.38) that people overall blamed action and inaction differently.

Next we tested the **penalty of dismissal** hypothesis: that deciders who dismiss the advisor's recommendation and decide the opposite would get blamed more strongly than deciders who followed the advisor's recommendation. There was decisive evidence in favor of the hypothesis (BF > 100). Dismissing deciders were blamed far more ($M$ = 57.6, $SD$ = 36.1) than complying deciders ($M$ = 40.9, $SD$ = 34.7); $d$ = 0.47. Notably, there was strong evidence that this dismissal penalty did not vary by the course of action the decider ended up taking (BF = 0.29). Instead, it varied by decider. Robot deciders were penalized even more strongly ($d$ = 0.67) than human deciders ($d$ = 0.30) for dismissing the advice given to them (BF = 8.2).

What is remarkable about the penalty of dismissal patterns is that people penalize *not following advice* more strongly than *not following the norm*. As seen in Fig. 3, the highest levels of blame apply to deciders who dismissed the advice given to them (red filled circles). By contrast, those who followed the advice were blamed much less (white-filled circles), and that was true even when they chose inaction—which goes against the norm in this scenario.

Finally, we examined the hypothesized **differential dismissal penalties** for *specific* combinations of human vs. robot advisors and human vs. robot deciders. We took the penalty for a human advisor-human decider pairing as the reference, which was 7.2 points—how much more human deciders were blamed when dismissing than accepting the human advice. We reasoned that this penalty would be larger when a robot decider dismisses a human advisor (because the advisor's greater moral expertise should be accounted for). There was decisive evidence that this penalty was indeed larger (17.1 points) when a robot decider dismissed a human advisor (BF > 100), $d$ = 0.36. Conversely, we reasoned that the penalty would be smaller when a human decider dismisses a robot advisor (because the decider's greater moral expertise grants justification for dismissing the advice). However, there was substantial evidence against the hypothesis of a lower penalty for the combination of robot advisor and human decider (BF = 0.24), $d$ = −0.12. As Fig. 4 illustrates, the Human → Robot pairing shows a much greater dismissal penalty than the other three.

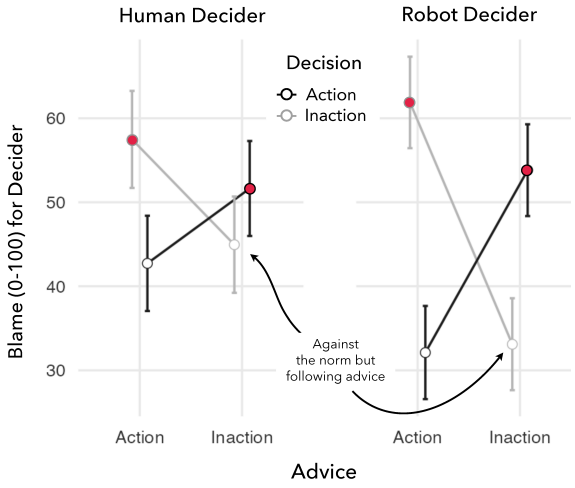


**Figure 3: Penalties of dismissal for a human decider (left panel) and a robot decider (right panel), depending on what advice was given (action vs. inaction) and what decision was made (black lines for action vs. grey lines for inaction). Red dots mark cases in which deciders dismissed the advice.**

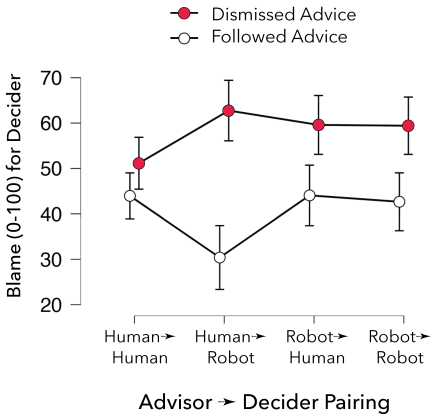


**Figure 4: Patterns of dismissal penalty across advisor-decider pairings: Mean blame, with 95% conf. intervals, for the decider (human vs. robot) depending on whether the decider dismissed or followed advice from a human vs. robot advisor.**

## 7 EXPERIMENT 2 DISCUSSION

Experiment 2 attempted to replicate the findings from Experiment 1 and extend them to people's integrated perception of (robot and human) moral advisors and decision makers.

As in Experiment 1, people applied the same norms to human and robot advisors in the presented dilemma. Next, we found again a strong disagreement effect: people blamed any advisor (human or robot) about 13 points more when the advisor recommended the opposite, rather than the same, course of action that participants endorsed. Finally, we replicated the finding (though with smaller effect size and confidence) that people treat advice similarly to

decision: they tend to blame a robot advisor more than a human when it recommends inaction, which is the kind of human-robot asymmetry repeatedly found for moral decisions [23, 31, 41].

Experiment 2 also allowed us to test several hypotheses about people's moral evaluations of the *decider*, in the context of being given advice. We found evidence against an inaction asymmetry: People did not blame humans less than robots for inaction decisions. The present context is not directly comparable to previous scenarios—in which people evaluated a single decision maker, without advisor—but the data do suggest some boundary conditions on the robustness of the inaction asymmetry for decisions.

Further, we found strong evidence for a penalty of dismissing advice: People blamed deciders on average 17 points more for dismissing the advice they received than for following it. This penalty varies, however, by agent type. Whereas human deciders suffered only a 7-point penalty when dismissing a human advisor, robot deciders suffered a 32-point penalty when dismissing a human advisor. Clearly, people expect robots to go along with human moral expertise. However, human deciders were also penalized when dismissing a robot's moral advice. The penalty was 15 points, and though it did not differ from the human-human baseline, this finding suggests that people regard even a robot's moral advice to a human decider as worthy of consideration.

## 8 GENERAL DISCUSSION

In two experiments, we investigated people's perceptions of humans and robots in moral advising scenarios. Our findings provide first insights into the general phenomenon of moral advising and the unique ways people assign blame to robot moral advisors.

### 8.1 Insights Gained

A first insight is that people consider moral advice not only as expressing the applicable norms but as indicating the advisor's moral decision making. As a result, advisors receive blame, in degree and pattern, similarly to how deciders received blame in previous studies. In both of our experiments, people blamed robot advisors more than human advisors for recommending inaction (i.e., refraining from sacrificing one person for the good of many). We can call this an inaction asymmetry in moral *advice*, whereas previously such an asymmetry was only seen in moral decisions [31, 32, 41].

Second, people's moral disagreement with an advisor's recommendation naturally led to substantial blame, but this disagreement effect was symmetric for humans and robots. This symmetry is consistent with the finding that people apply the same norms to human and robot agents in moral dilemmas [23, 33, 41]. Likewise, our results do not suggest across-the-board higher blame for robots than for humans. Even though, as Fig. 2 shows, robot advice tended to be blamed more than human advice in Experiment 1, this pattern did not emerge in Experiment 2, nor was there overall higher blame for robot deciders in Experiment 2. Moreover, this lack of greater overall blame for robots occurred despite our exclusion of a number of people for disqualifying robots as potential targets of blame. When people disqualify robots in this way, they typically give low or 0 blame ratings; thus we removed people who would have brought *down* average robot blame. Even after this correction, blame for robots did not exceed blame for humans. We conclude that, when scenarios are sufficiently detailed and robots are described as having important capacities, people apply the same norms to humans and robots and blame them to similar degrees [12, 53]). The human-robot asymmetry for inaction may be an exception, stemming from spontaneous and brittle empathy with reluctant human deciders.

A third insight is that the presence of advisors has a notable impact on how people morally evaluate deciders. When we tested the inaction asymmetry on the human vs. robot *deciders* in Experiment 2, the familiar human-robot asymmetry failed to emerge. But unlike the deciders in previous studies, deciders in this experiment had to make their decision in light of an assistant's advice. So our results suggest that in the presence of an advisor, human deciders are no longer spared blame for refraining from sacrificing one for the good of many. Whereas people seem to understand why a human decision maker refrains from acting when caught alone in a dilemma, this understanding does not extend to a decision maker who is being advised.

The fourth insight is that the obligation to follow advice was stronger than the obligation to make the norm-prescribed decision (which is, in the present scenario, to act). As Fig. 3 shows, people strongly blamed deciders who dismissed the advice given to them, even when that advice promoted the normatively disfavored choice—and perhaps *should have* been dismissed. Thus, people saw advice to not act (even though it is against the norm) as an obligation for deciders to follow the advice, not the norm.

Fifth, the force of advisors, and hence the penalty for dismissing advice, varied across pairings of human vs. robot advisors and deciders. Relative to a small human → human penalty, the human → robot penalty was substantial, suggesting that people most firmly objected to a robot that dismissed human advice and most strongly mitigated blame for the robot that followed human advice (see Fig. 4). While robots may at times need to ignore or disobey inappropriate human commands [7, 25], in the present context, participants considered the human advice to the robot nearly binding. Interestingly, the other two conditions (in which a robot gave advice) elicited dismissal penalties as well, no less than the baseline human-human penalty. Participants took the robot advisor seriously, and expected human and robot deciders to do so as well. Thus, we see again that, when robots are described as capable of making moral recommendations, people may readily value those recommendations; a finding that raises unique opportunities [cf. 54] but also acute concerns [cf. 17]. This stands in contrast to other studies, where sparse descriptions of robots that had few capabilities made people reluctant to embrace a robot as a moral decider [5].

### 8.2 Broader Implications

Our results have intriguing implications for the design of morally competent and language-capable robots, as they suggest that if robots are designed to be advisors in morally significant situations, their recommendations must strongly take into account the common good. By contrast, human moral advisors are blamed less for refusing to sacrifice one individual to serve the common good. Thus, whereas human advisors are partially forgiven, perhaps out of empathy, for refusing to serve the common good, robots are expected to more consistently adhere to these norms and will not escape blame for recommending inaction.

Yet this does not suggest that people want robots to be "utilitarians." People clearly express that the norm for *both* humans and robots is to to take action in this situation [23, 31, 32]. While people seem to make allowance for a human's difficulty to sacrifice another human [41], this may be an empathic response rather than reveal any broader inclinations toward or away from utilitarianism.

When we examine how people morally evaluate *deciders* (Experiment 2), the pattern of results was different, adding to the advice/choice asymmetries in other literatures [2, 36]. Specifically, despite consistent evidence in past research for an inaction asymmetry for human vs. robot deciders, Experiment 2 did not find such an asymmetry. The key difference in our experiment is that human deciders make their choice under another agent's *advisement*. In this case, they seem no longer forgiven for choosing inaction, and the previously found human-robot inaction asymmetry disappeared.

The impact of moral advice on human judgments was substantial. People expected both robot and human deciders to follow their advisor's recommendation even when the advisor recommended to violate the norm-prescribed choice. Because this is the first time such a pattern has been found, we must be cautious in weighting it too heavily; but the implications are noteworthy for situations of hierarchy and strong social bonds: Advice from either a human or a robot to perform a morally unpopular action may be sufficient to force a decider to take that action. Because this is a moral dilemma, even the unpopular action is defensible, but it still goes against what three-fourths of our participants said the agent should do [23, 41].

Robots, especially, received more blame when failing to follow advice, suggesting that observers expect them to consistently adhere to advice, especially human advice (see Fig. 4). This could stem from perceiving robots as less capable of making difficult moral decisions. Or, observers might believe robots are capable but shouldn't have the autonomy to disobey commands in such situations.

## 8.3 Limitations

Several limitations must be considered when interpreting our results. First, the study's online nature presented the scenario to observers in a text-based, image-accompanied format. In an in-person setting, observers might have reacted differently, and their judgments could have been influenced by additional contextual cues. The real world, of course, also introduces numerous uncontrolled (or unknown) factors, including the physical features of a real robot, the physical presence of an experimenter, and so forth. Moreover, no in-person study could reasonably run 1000 participants. But now that we see the unique and powerful features of advisor-decider interactions, we can plan more focused studies (e.g., specific cells from the large design) that are feasible to conduct in the lab.

Second, participants were observers, not advisors or deciders. The ethical challenges of placing participants in scenarios resembling the trolley problem are substantial, even in virtual reality. Future studies could attempt to immerse participants in a dialog context in which they role-play an advisor or decider in a remote-communication setting and make choices under time pressure.

Third, participants were always asked to choose between two types of advice, rather than whether or not to advise. Future work could further consider the effects of advisors who are able to give advice but who choose not to provide advice on a dilemma.

Fourth, our data analysis focused primarily on quantitative data and, even though we did collect qualitative responses from participants, we did not use these data for the main results. A deeper exploration of these qualitative responses could potentially offer valuable insights into the diverse ways observers make blame judgments. We did use the qualitative responses to identify participants who rejected robots as targets of blame and to remove them from the analysis. Further explorations of these rejections could be valuable: For example, participants might refuse to blame robots in the context of moral dilemmas but perhaps not in the context of medical or financial advice.

Finally, our experiment offered depictions of a mechanomorphic robot, which is the baseline robot agent in previous moral HRI work [32]. However, a different (e.g., more humanoid) robot design might alter observers' moral judgments [32]. Future research could benefit from replicating this study with a broader range of robot types, which could offer a deeper understanding of how different robot designs affect observers' perceptions and judgments.

## 8.4 Open Questions

This experiment raises challenging questions. How should humans uptake and use moral advice? Are they really expected to follow advice (even robots' advice) at the expense of following norms? Our results also raise questions about the design of morally competent robots. Should such 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)? Should such morally competent robots always prioritize actions that minimize blame, even if this avoidance of blame leads them to follow morally questionable advice? Balancing the ethical requirements of advice-following and moral correctness poses complex challenges to robot design.

These questions underscore the intricacies of human-robot interactions in moral scenarios, and the need for ethical discussions and careful consideration when developing and deploying robots.

# 9 CONCLUSION

We examined human perceptions of moral advising behaviors in a moral dilemma. Our results suggest that even when a robot is advising a human in a moral dilemma, people expect the robot to adhere to moral principles as if it were making the decision itself. We further discovered that moral advice may enact a strong obligation on decision makers receiving advice. People's moral blame for dismissing advice was substantial, even when the advice was the morally unpopular course of action. Moral decision makers may find themselves in a second dilemma: avoid blame by following the general norm in the situation, or avoid blame by following the advice—no matter whether it expresses the general norm or not.

# ACKNOWLEDGMENTS

Tom Williams' work was funded in part by National Science Foundation grant, and by Air Force Office of Scientific Research (AFOSR) Young Investigator Award 19RT0497. Bertram Malle's work was funded in part by AFOSR Award FA9550-21-1-0359. The authors would like to thank Katherine Aubert for her assistance.

# REFERENCES


[1] Florian Aschauer, Matthias Sohn, and Bernhard Hirsch. 2023. Managerial advice-taking: Sharing responsibility with (non)human advisors trumps decision accuracy. *European Management Review* (2023). https://doi.org/10.1111/emre.12575

[2] Rachel Barkan, Shai Danziger, and Yaniv Shani. 2016. Do as I say, not as I do: Choice–advice differences in decisions to learn information. *Journal of Economic Behavior & Organization* 125 (May 2016), 57–66. https://doi.org/10.1016/j.jebo.2016.02.005

[3] Meir Barneron and Ilan Yaniv. 2020. Advice-giving under conflict of interest: Context enhances self-serving behavior. *Journal of Experimental Social Psychology* 91 (Nov. 2020). https://doi.org/10.1016/j.jesp.2020.104046

[4] Daniel Ben-David, Yehezkel S. Resheff, and Talia Tron. 2021. Explainable AI and adoption of financial algorithmic advisors: An experimental study. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 390–400. https://doi.org/10.1145/3461702.3462565

[5] Yochanan E. Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (Dec. 2018), 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

[6] Gordon Briggs. 2014. Blame, What is it Good For?. In *Proceedings of the Workshop on Philosophical Perspectives on HRI at RO-MAN 2014*. Edinburgh, Scotland.

[7] Gordon Briggs, Tom Williams, Ryan Blake Jackson, and Matthias Scheutz. 2022. Why and how robots should say 'no'. *International Journal of Social Robotics* 14, 2 (March 2022), 323–339. https://doi.org/10.1007/s12369-021-00780-y

[8] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the International conference on Human-Robot Interaction (HRI)*. ACM, 293–300.

[9] Jason Dana and Daylian M Cain. 2015. Advice versus choice. *Current Opinion in Psychology* 6 (Dec. 2015), 173–176. https://doi.org/10.1016/j.copsyc.2015.08.019

[10] Philippa Foot. 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review* 5 (1967), 5–15.

[11] Russell Fulmer, Tonya Davis, Cori Costello, and Angela Joerin. 2021. The ethics of psychological artificial intelligence: Clinical considerations. *Counseling and Values* 66, 2 (Oct. 2021), 131–144. https://doi.org/10.1002/cvj.12153

[12] Caleb Furlough, Thomas Stokes, and Douglas J. Gillan. 2021. Attributing blame to robots: I. The influence of robot autonomy. *Human Factors* 63, 4 (June 2021), 592–602. https://doi.org/10.1177/0018720819880641

[13] Alberto Giubilini and Julian Savulescu. 2018. The artificial moral advisor: The "ideal observer" meets artificial intelligence. *Philosophy & Technology* 31, 2 (June 2018), 169–188. https://doi.org/10.1007/s13347-017-0285-z

[14] Liat Hadar and Ilan Fischer. 2008. Giving advice under uncertainty: What you do, what you should do, and what others think you do. *Journal of Economic Psychology* 29, 5 (Nov. 2008), 667–683. https://doi.org/10.1016/j.joep.2007.12.007

[15] Michael C Horowitz and Erik Lin-Greenberg. 2022. Algorithms and influence artificial intelligence and crisis decision-making. *International Studies Quarterly* 66, 4 (Dec. 2022). https://doi.org/10.1093/isq/sqac069

[16] Ryan Blake Jackson and Tom Williams. 2018. Robot: Asker of questions and changer of norms? *Proceedings of ICRES* (2018).

[17] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In *Companion Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*. IEEE, 401–410.

[18] 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*.

[19] Harold Jeffreys. 1948. *Theory of probability.* (2d ed. ed.). Clarendon Press, Oxford.

[20] James Kennedy, Paul Baxter, and Tony Belpaeme. 2014. Children comply with a robot's indirect requests. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot interaction (HRI)*. 198–199.

[21] 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. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, USA, 10–18. https://doi.org/10.1145/3434074.3446908

[22] Michal Klincewicz. 2017. Artificial intelligence as a means to moral enhancement. *Studies in Logic, Grammar and Rhetoric* 48, 1 (Dec. 2017), 171–187. https://doi.org/10.1515/slgr-2016-0061

[23] 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.

[24] Laura Kray and Richard Gonzalez. 1999. Differential weighting in choice versus advice: I'll do this, you do that. *Journal of Behavioral Decision Making* 12, 3 (1999), 207–218. https://doi.org/10.1002/(SICI)1099-0771(199909)12:3<207::AID-BDM322>3.0.CO;2-P

[25] Michael Laakasuo, Jussi Palomäki, Anton Kunnari, Sanna Rauhala, Marianna Drosinou, Juho Halonen, Noora Lehtonen, Mika Koverola, Marko Repo, Jukka Sundvall, Aku Visala, and Kathryn B. Francis. 2023. Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology* 53, 1 (2023). https://doi.org/10.1002/ejsp.2890

[26] Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course.* Cambridge university press.

[27] 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*. 695–704.

[28] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human perceptions on moral responsibility of AI: A case study in ai-assisted bail decision-making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3411764.3445260

[29] Bertram F Malle. 2016. Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots. *Ethics and Information Technoogy* 18, 4 (2016), 243–256.

[30] Bertram F Malle and Matthias Scheutz. 2014. Moral Competence in Social Robots. In *Symposium on Ethics in Science, Technology and Engineering*. IEEE, 1–6.

[31] 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.

[32] 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 *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 125–132.

[33] 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*, Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi, and Endre E. Kadar (Eds.). Springer International Publishing, Cham, 111–133. https://doi.org/10.1007/978-3-030-12524-0_11

[34] Terran Mott and Tom Williams. 2023. Confrontation and Cultivation: Understanding Perspectives on Robot Responses to Norm Violations. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2336–2343.

[35] Mauricio Palmeira, Gerri Spassova, and Hean Tat Keh. 2015. Other-serving bias in advice-taking: When advisors receive more credit than blame. *Organizational Behavior and Human Decision Processes* 130 (Sept. 2015), 13–25. https://doi.org/10.1016/j.obhdp.2015.06.001

[36] Evan Polman and Rachel L. Ruttan. 2022. Making utilitarian choices but giving deontological advice. *Journal of Experimental Psychology: General* 151, 10 (Oct. 2022), 2614–2621. https://doi.org/10.1037/xge0001194

[37] Christina A. Rader, Richard P. Larrick, and Jack B. Soll. 2017. Advice as a form of social influence: Informational motives and the consequences for accuracy. *Social and Personality Psychology Compass* 11, 8 (Aug. 2017), e12329. https://doi.org/10.1111/spc3.12329

[38] 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 International Conference on Human-Robot Interaction (alt.HRI)*. 21–29.

[39] Blanca Rodríguez-López and Jon Rueda. 2023. Artificial moral experts: Asking for ethical advice to artificial intelligent assistants. *AI and Ethics* 3 (Jan. 2023), 1371–1379. https://doi.org/10.1007/s43681-022-00246-5

[40] 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 *Proceedings of the International Conference on Human Robot Interaction (HRI)*. IEEE Press, 117–124.

[41] 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*, J. Gaillot, Derek Macintosh, and J. D. Ohlin (Eds.). Oxford University Press, Oxford, UK, 89–102.

[42] 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)*. ACM, 479–486.

[43] 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. https://doi.org/10.1007/978-3-030-62056-1_23

[44] 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 *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 178–186.

[45] Marco Tassella, Rémy Chaput, and Mathieu Guillermin. 2023. Artificial moral advisors: Enhancing human ethical decision-making. In *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*. 1–5. https://doi.org/10.1109/ETHICS57328.2023.10155026

[46] Don van den Bergh, Johnny van Doorn, Maarten Marsman, Tim Draws, Erik-Jan van Kesteren, Koen Derks, Fabian Dablander, Quentin F. Gronau, Šimon Kucharský, Akash R. Komarlu Narendra Gupta, Alexandra Sarafoglou, Jan G. Voelkel, Angelika Stefan, Alexander Ly, Max Hinne, Dora Matzke, and Eric-Jan Wagenmakers. 2020. A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année psychologique* 120, 1 (2020), 73–96. https://doi.org/10.3917/anpsy1.201.0073

[47] Ruchen Wen, Zhao Han, and Tom Williams. 2022. Teacher, teammate, subordinate, friend: Generating norm violation responses grounded in role-based relational norms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 353–362.

[48] 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.

[49] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Brščić, Iolanda Leite, and Tom Williams. 2022. Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 120–129.

[50] Katie Winkle, Séverin Lemaignan, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2019. Effective persuasion strategies for socially assistive robots. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 277–285.

[51] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction*. 29–37.

[52] Yuyan Xia, Yanchun Chen, Huiyuan Luo, Yuer Yang, and Xiangjie Wang. 2022. Research and improvement on the development of robo-advisor: present and prospect. In *Proceedings of the 2022 4th International Conference on Image, Video and Signal Processing (IVSP '22)*. Association for Computing Machinery, New York, NY, USA, 172–178. https://doi.org/10.1145/3531232.3531257

[53] April D. Young and Andrew E. Monroe. 2019. Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology* 85 (Nov. 2019), 103870. https://doi.org/10.1016/j.jesp.2019.103870

[54] 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.