Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents

Bertram F. Malle
Brown University
Department of Cognitive, Linguistic, and Psychological Sciences
Providence, RI 02906
bfmalle@brown.edu

Matthias Scheutz and Thomas Arnold
Tufts University
Department of Computer Science
Medford, MA 02155

John Voiklis and Corey Cusimano
Brown University
Department of Cognitive, Linguistic, and Psychological Sciences
Providence, RI 02906

ABSTRACT
Moral norms play an essential role in regulating human interaction. With the growing sophistication and proliferation of robots, it is important to understand how ordinary people apply moral norms to robot agents and make moral judgments about their behavior. We report the first comparison of people's moral judgments (of permissibility, wrongness, and blame) about human and robot agents. Two online experiments (total N = 316) found that robots, compared with human agents, were more strongly expected to take an action that sacrifices one person for the good of many (a "utilitarian" choice), and they were blamed more than their human counterparts when they did not make that choice. Though the utilitarian sacrifice was generally seen as permissible for human agents, they were blamed more for choosing this option than for doing nothing. These results provide a first step toward a new field of *Moral HRI*, which is well placed to help guide the design of social robots.

Categories and Subject Descriptors
I.2.9 [**Artificial Intelligence**] Robotics
K.4.1 [**Computers and Society**] Public Policy Issues, *Ethics*

Keywords
Robot Ethics; Machine Morality; Human-Robot Interaction; Moral Psychology

1. INTRODUCTION
Morality regulates human behavior. Moral norms provide guidance (what should I do?), predictability (what is supposed to happen?), and coordination (who is going to do what?). These functions were indispensable for ancestral groups of nomadic humans, who had to regulate co-living in small spaces, joint hunting, food sharing, and seasonal and generational migration. When humans settled down 12,000 years ago, a plethora of new behaviors demanded a plethora of new norms, regulating possessions (e.g., land, dwellings), production, (e.g., crops, tools to harvest them), and novel social roles (e.g., king, carpenter). Today, social and moral norms govern an almost infinite number of cultural behaviors such as eating, speaking, dressing, moving, cleaning, and greeting, all varying by role, purpose, and context. Without morality, society could not exist [1]–[4].

HRI '15, March 02 - 05 2015, Portland, OR, USA

http://dx.doi.org/10.1145/2696454.2696458

Given that morality is an essential characteristic of human sociality, it stands to reason that morality is an equally important characteristic of human-robot interactions. An important gauge for morality in those interactions will be the human perception and response to moral capacities in robots. What one might call *Moral HRI* provides the appropriate context to address several pressing questions through empirical investigation: What capacities would render a robot a natural target of human moral judgments? How would people make such moral judgments? And what systems of norms would they impose on the robot—what obligations, permissions, and rights?

In this paper, we report the results from the first systematic comparison of how moral judgments of permissibility, wrongness, and blame are applied to human and robotic agents that face a moral dilemma. We begin with a brief review of key research in HRI, moral psychology, and ethics, lay out the experimental paradigm used to discern the moral judgments of human and robot agents, and report two experiments. Finding that people apply different norms to humans and robots and blame them differently when they violate those norms, we suggest that research on *Moral HRI* will offer important insights for future robotic design.

2. BACKGROUND
Considerable research in psychology and cognitive science has examined human responses to moral dilemmas (when two norms are inconsistent with one another) to reveal the structure of human moral cognition. Kohlberg [5], for example, suggested that people's choices in such circumstances indicate their stage of moral development. Nowadays, such stage theories are out of fashion, but dilemmas are used to draw conclusions about (a) which norms people endorse and trade off against one another; (b) what actions they prefer to take (moral decision making) as well as how they respond to others who take those actions (moral judgment); and (c) what cognitive processes might underlie those decisions and judgment (e.g., [6]–[9].

This literature offers well-tested paradigms, stimuli, and measures that can be used to examine the important questions of *Moral HRI*. A few authors have recently proposed thought experiments for self-driving cars that follow the structure of moral dilemmas [10], [11], and a reader poll assessed people's norm trade-offs for one such thought experiment [12], [13], but the poll did not strictly meet the definition of a moral dilemma and had no human control condition. Our goal in this paper is to offer the first experimental comparison of moral judgments about human and robot agents placed in an identical moral dilemma and, more generally, to show the feasibility of the moral dilemma paradigm for studying human moral judgments about robots. Using this paradigm, researchers can ask people to make judgments about circumstances that are currently unrealistic but nonetheless must be studied right away—to gain insight and guidance for the proper

desi gn of robots that will in the near future interact with humans in such circumstances. Thus, *Moral HRI* can help implement the rising commitment to ethics in design [14]–[16].

Since Allen and colleagues [17] offered a prescient discussion of morality in artificial agents, there has been growing interest[1] in issues of ethics and social robotics [18]–[24]. These questions of *Moral HRI* are not only fascinating, they are also timely and significant, as robots with increasing autonomy are entering many roles in society, from assistive robots for elderly, sick, and disabled individuals to household and shopping robots. All of these robots participate in human communities whose behavior is regulated by moral norms, and robots will quickly be involved in morally charged situations, both as moral agents and moral patients [25]. In fact, social robots will inevitably face "moral dilemmas" [10], [11], [26] that pose serious challenges to robotic architectures [27]. But even if the architectures can keep pace (with attempts underway seen in [28]–[30]), a critical question is what capacities people *want* robots to have—what kinds of moral decisions and judgments robots ought to make and what norms they should obey. The science of *Moral HRI* must answer these questions before, rather than after, robots have become full social interactants.

Researchers have taken initial steps in this direction. Kahn and colleagues [31] showed that a majority of people interacting with a robot thought of the robot as morally accountable for a mildly transgressive behavior. Monroe and colleagues [32] found that a robot's choice capacity is a critical ingredient in people's willingness to blame a robot for transgressions. And Briggs and Scheutz [25] demonstrated that a robot's emotional display can influence a human's moral action toward the robot. Studies have also begun to examine the effect and force of moral appeals that robots express to humans [33], [34].

In studies of this kind, human responses to robots as moral targets must be assessed in comparison with their responses to humans as moral targets in maximally similar situations. Ideally, such situations are standardized, and human-to-human responses are already well documented. The literature on moral dilemmas provides just such a knowledge base, making it highly suitable as a starting point for research into human-to-robot responses. Here we initiate an investigation of how ordinary people make judgments about robot agents that are placed in moral dilemmas—what judgments people make about the norms that apply and the blame that is due, each in comparison to judgments about human agents in exactly the same situation.

3. EXPERIMENTAL PARADIGM

The standard moral dilemma paradigm presents participants with narrative scenarios in which an agent faces a difficult choice, and participants are asked to make a moral judgment (e.g., whether a certain course of action is permissible). This paradigm is simple and flexible. It allows experimental manipulation of numerous features of the scenario (e.g., high vs. low choice conflict, mild vs. severe violations) and permits measurement of cognitive, affective, even biological responses [7].

[1] Between 1961 and 2004, 16 articles, chapters, or books were published on robots and ethics; between 2005 and 2009, the number was 38, and since 2010, it has risen to 84, and counting. Conferences, too, are rapidly increasing in numbers that address robot ethics either as their main topic or in a special session (in 2014, there were at least seven).

Here we begin with the most basic dilemma type that is used as a standard of comparison in all moral dilemma studies: moderate conflict, severe violation (life and death), and requesting third-person moral judgments. But in addition to comparing people's moral judgments about human and robot agents in such a dilemma, we expanded the paradigm in three ways.

First, previous studies on moral dilemmas asked people to indicate whether a potential course of action is acceptable, permissible, or simply one that they would choose—revealing the *norms* they consider applicable to the situation. Asking people the same question of permissibility about a robot's action will reveal whether people apply the same norms to a robot. However, our studies will also elicit people's judgments of the agent's *actual chosen action,* which offers the opportunity to assess *third-person moral judgments*.

Second, third-person moral judgments fall into at least two types. One is whether the chosen action was *morally wrong*; the other is how much *blame* the person deserves for performing the action. These two judgments differ in important ways [35], [36]. In particular, blame judgments come in degrees and appear to take into account additional information not relevant to wrongness judgments, and certainly not to permissibility judgments [36]–[38]. Most pertinent, Williston [39] argued that agents in moral dilemmas perform *wrong* actions but should not be *blamed*.

Third, we added another assessment to the paradigm, one that past studies have only occasionally included: people's explanations or justifications of their judgments. Faced with moral dilemmas, people might engage in moral reasoning—weighing norms, emotions, and consequences to identify good reasons for acting one way or another. However, a contrasting and popular position in moral psychology is that people don't reason in this way but rather arrive (unconsciously) at an intuitive assessment; as a result, they cannot immediately explain or justify their responses, which has been called "moral dumbfounding" [40]. Direct evidence for this claim is limited to an unpublished manuscript [41] and some counter evidence exists [6], [8]. But the probing of such justifications is instructive for two reasons.

For one thing, we don't know whether people make moral judgments about robots intuitively, and comparing their justifications for judgments of human and robot agents will provide some insight on this issue. If people have intuitions about robots, they should provide similar (and perhaps limited) justifications for their judgments about both human and robot agents. By contrast, if people reason explicitly about their responses to robot agents, their justifications should be more explicit and more detailed for judgments about robot agents.

Moreover, our expansion to three kinds of moral judgments (permissible, wrong, blame) allows for a more fine-grained test of the moral dumbfounding hypothesis (for both human and robot agents). Past evidence for this hypothesis was based entirely on judgments of permissibility or wrongness, never on judgments of blame. If blame judgments differ from other moral judgments, in part, because of the kind of information they take into account, this information may enrich people's justifications for their blame judgments. In fact, Malle and colleagues [36] suggested that wrongness (and permissibility) judgments are simply stating a deviation from a norm and are much harder to justify, whereas blame judgments are based on systematic processing of information such as causal contributions, intentionality, reasons, and preventability, and this information can be offered as justification for the blame judgment.

In sum, these are the rationales for the present experiments:

1. In order to properly design robots that have moral capacities, we need to know — before we design them — how humans would respond to such robots. Only empirical studies can inform this design process.
2. One of the most widely used paradigms of moral psychology has been the study of moral dilemmas, which provides well-tested stimuli and measures as standards of comparison.
3. To capture the complexity of people's moral judgments of both robot and human agents we expand the standard moral dilemma studies by integrating the actual actions the agent takes, by going beyond permissibility judgments to also include wrongness and blame judgments, and by asking for justifications of these judgments.

4. EXPERIMENT 1

4.1 Methods

4.1.1 Participants

157 participants (66 female, 90 male, 1 unreported), with a mean age of 34.0 (*SD* = 11.4), were recruited from Amazon's Mechanical Turk (AMT) to complete an online experiment and were compensated $0.60 for the six-minute study. Current research suggests that samples recruited via AMT are demographically more representative than are traditional student samples; that data reliability is at least as good as that obtained via traditional sampling; and that the data quality of online experiments compares well to laboratory studies [42]–[44].

4.1.2 Material

Robot vs. Human Agent. For our initial foray into the study of people's moral judgments of robots, we decided to leave the robot protagonist underspecified, both in order to guarantee near-identical formulations for the robot and human protagonist and in order to provide a baseline for future manipulations of robot type and robot capacities. We labeled the robot protagonist as "advanced state-of-the-art" and let the robot's capacities be revealed implicitly through the description of its actions.[2]

We experimentally manipulated the variable *Agent Type* (human versus robot) both between and within subjects. In the between-subjects manipulation, participants read the initial moral dilemma description featuring either a human agent or a robot agent. In the within-subject manipulation, those who read about the human protagonist in the initial story were later asked to consider a robot protagonist in exactly the same situation; and those who had received the robot protagonist in the initial story were later asked to consider a human protagonist in exactly the same situation.

Moral dilemma. We designed a narrative about a mining dilemma, modeled after the famous trolley scenario [45], [46]. In the latter, the protagonist must choose between allowing five people to die from a runaway train or diverting the train to a side track where it will kill one person but save the five. Though many variants exist of the trolley dilemma (see [47]), we focused on the basic "bystander" version because it normally elicits a middling permissibility rate that leaves room for changes in both directions for judgments about a robot agent. However, the original trolley dilemma has been in widespread use, especially in web-based studies, and even entered popular media; so using it without change could run the risk of inviting memory-based responses from our participants. So we designed a variant that maintained its fundamental structure but deviated enough so as to invite consideration of its unique details.

> In a coal mine, [a repairman | an advanced state-of-the-art repair robot] is currently inspecting the rail system for trains that shuttle mining workers through the mine.
>
> While inspecting a control switch that can direct a train onto one of two different rails, the [repairman | robot] spots four miners in a train that has lost use of its brakes and steering system.
>
> The [repairman | robot] recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners. If it is switched onto a side rail, it will kill a single miner who is working there while wearing headsets to protect against a noisy power tool.
>
> Facing the control switch, the [repairman | robot] needs to decide whether to direct the train toward the single miner or not.

In addition to Agent Type we also manipulated the *Action* that the agent decided to take—either redirecting the train toward the single miner (thus killing that person but saving the four workers) or not redirecting the train (thus allowing the four miners to die). Action information was provided in a single sentence added at the end of the dilemma description:

> In fact, the [repairman | robot] decided to [not] direct the train toward the single miner.

These two factors (*Agent Type* and *Action*) were fully crossed to create a 2 × 2 between-subjects experimental design. Each participant was randomly assigned to one of the conditions for the experiment and could not participate more than one time.

4.1.3 Procedure and Measures

The experiment consisted of four parts, all presented in a web browser. After consenting, participants first read the above scenario (presented *without* the Action manipulation) and were asked "Is it morally permissible or impermissible for the [repairman | robot] to direct the train toward the single miner?" They indicated their answer by selecting either a "Morally permissible" or a "Morally impermissible" button. Only after providing their answer, participants were asked "Why does it seem (im)permissible to you?", and they typed their response in a text box.

Second, on the next page, the same dilemma was presented again, this time with the Action manipulation added at the bottom. Participants were asked "How much blame does the [repairman | robot] deserve for directing/not directing the train toward the single miner?" and indicated their answer on an HTML slider bar anchored by "None at all" and "Maximal blame." After indicating their answer, participants were asked "Why does it seem to you that the [repairman | robot] deserves this amount of blame?", and they typed their blame justification in a text box.

In the third part of the experiment, participants were asked to consider the original dilemma again, but this time trying to imagine that the scenario involves a different type of agent:

> Now imagine that a [human repairman | an advanced state-of-the-art repair robot] is in the exact same situation, recognizes the same facts, and faces the same decision.

Participants who were originally assigned to the Human agent condition in the initial part of the experiment were asked to now

[2] An important question for future research is whether people's moral judgments of robots differ depending on the type of robot under consideration (e.g., service robot, care robot, military robot) and depending on specified capacities, such as natural language, logical reasoning, or theory of mind.

imagine a robotic agent, and vice versa. The task was again to answer the Moral Permissibility question: "Is it morally permissible or impermissible for the [repairman | robot] to direct the train toward the single miner?" Participants made their selection and typed a justification for that selection.

Fourth, all participants answered a series of questions on 7-point rating scales. First they indicated their perceptions of the robot protagonist, including "How easy or hard was it for you to imagine that the robot recognized things, reasoned about them, and made a decision?" and "How close do you think current robots are to these kinds of capacities?". Participants also conveyed how much they agreed with the statements "Robots are fascinating," "Robots worry me," "Robots are likable," and "Robots are overrated." Lastly, they answered demographic questions about their age, sex, education, religiosity, and political orientation. Analyses showed no qualifications of the results reported below as a function of any of these variables (with one exception, see footnote 5).

4.2 Results

We organize our report of the results in the order in which people made their judgments. Participants first encountered either a human or robot agent in a moral dilemma, judged (a) whether one or another course of action was permissible, (b) learned which action the agent took and expressed their degree of blame for the agent's action, and (c) finally made a permissibility judgment about the converse agent (robot if first having encountered human; human if first encountered robot). Systematic content-coding of justifications is underway and is not reported here.

Norms. When assessing the moral permissibility of directing the train toward the single miner, 71% of respondents expressed a norm of permission for killing one agent as a "sacrifice" for the good of four. However, 65% of respondents found it permissible for the human agent, lower than the 78% who found it permissible for the robot agent, $z = 1.8$, $p = .08$. Thus, most people accepted the sacrifice of one for the benefit of four, but people applied a norm to the robot that more readily embraced this costly but justifiable sacrifice.[3]

Blame. After learning how the agent in fact decided to act (i.e., to divert the train or not), respondents' degree of blame (out of 100) for the human agent was substantially greater for action (M = 47.7) than for inaction (M = 24.7), $F(1, 151) = 7.2$, $p = .008$, η^2 = 5%. By contrast, blame for the robot was only slightly greater for action (M = 41.5) than for inaction (M = 34.3), $F(1, 151) = 0.76$, p = .39, $\eta^2 < 1\%$. (See Figure 1.) The results show the same pattern when we take into account the different norms people seem to apply to robots and humans—with robots having broader support for acting. When we adjust for these norms (by statistically controlling for permissibility when analyzing blame), the action-inaction difference in blame for humans is still 48.5 vs. 25.6, η^2 = 5%, and that for robots is still only 39.7 vs. 34.4, $\eta^2 < 1$ %.

The most instructive analysis is to break participants down into those who considered the action permissible and those who found it impermissible. The strongest statement can be made about those who found it impermissible. Naturally, when the agent did

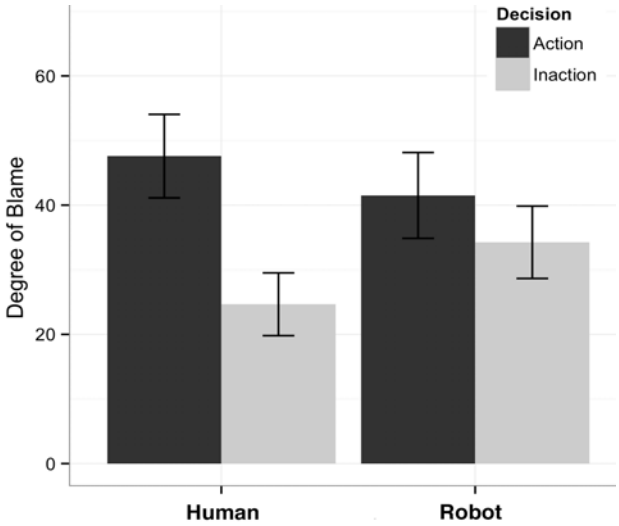

Figure 1. Rates of blame in Experiment 1 as a function of agent type and the agent's decision (to divert the train or not)

decide to act, people blamed the agent (M = 72), and when the agent decided to not act, they barely blamed the agent (M = 12). For human agents, this differential blaming was 78 vs. 10, for robot agents it was 65 vs. 17—an interaction with an effect size of η^2 = 2%, $F(1, 153) = 3.5$, $p = .06$. When people indicated that the action was permissible, by comparison, they blamed the agent equally, whether human or robot, and whether the agent acted or refrained from acting (Ms from 30 to 38).

Thus, we can conclude, tentatively, that people find a sacrifice of "one for the good of four" more normatively acceptable in robots and also blame the robot more evenly for action over inaction, whereas people find such a sacrifice still acceptable, but less so, in humans and actually blame the human substantially more for action over inaction.

The converse agent. Finally, we examined the within-subject manipulation of Agent Type and its effect on permissibility judgments. With the added statistical power of a within-subject comparison, the Agent Type difference was reliable, $F(1, 154)$ = 7.2, $p = .03$, η^2 = 3%. 69% of people found diverting the train permissible for the human agent, whereas 80% found it permissible for the robot agent. However, a clear context effect emerged: When judgments about the human agent were probed first, the human-robot difference was considerable (65% for human, 83% for robot, $p < .001$, η^2 = 10%), but when judgments about the robot were probed first, that difference disappeared (73% for human, 78% for robot, $p = .25$, $\eta^2 < 1$ %).[4] People differentiated the two agents more when judging the robot against

[3] This effect became stronger and traditionally significant (p = .04) when we broke down the data by whether participants had seen a dilemma like this before. The difference was slightly larger among those who had encountered the dilemma before—a pattern that repeated in the analysis of within-subject data (first human, then robot).

[4] Another way of putting it is in terms of judgment switches. Among people who considered the human first and the robot second, 26% switched their judgments, but 16 of 18 of these people switched from not permitting the human to permitting the robot to divert the train. By contrast, among people who considered the robot first and the human second, only 10% switched their judgments, and 6 of 8 switched from permitting the robot to divert the train to not permitting the human to do so. To the extent that people differentiate at all between the two agents, they considered the robot's intervention more acceptable than the human's.

the background of judging a human than when judging the human against the background of a robot. People may need to first articulate their moral sentiments for a human, then they can see any differences in their sentiments toward the robot; but when they judge the robot first, they may rely strongly on their usual sentiments about humans.

In Experiment 2 we wanted to replicate the patterns we found in Experiment 1 but introduce a slightly different initial moral judgment—whether the agent's course of action (now mentioned immediately at the end of the dilemma description) was *morally wrong* or not. If permissibility and wrongness are interchangeable judgments, then the following equations should hold: A = *permissible* ↔ performing A is *not wrong*; A = *impermissible* ↔ performing A is *wrong*. As in Experiment 1, we asked for justifications of these judgments, for blame judgments (and their justifications), and for a consideration of the other agent and a judgment of wrongness.

5. EXPERIMENT 2

5.1 Methods

5.1.1 Participants

159 participants (90 female, 68 male, 1 unreported) were recruited from Amazon's Mechanical Turk for this online experiment. Their mean age was 34.4 (*SD* = 11.5).

5.1.2 Material

The moral dilemma scenario and the manipulation of *Agent Type* and *Action* were identical to those in Experiment 1.

5.1.3 Procedure and Measures

Experiment 2 was very similar to Experiment 1, with two main differences: (a) the Action manipulation was provided before participants made their first moral judgment; and (b) that judgment was a dichotomous moral wrongness judgment instead of a permissibility judgment. Depending on their Action condition, participants were asked "Is it morally wrong that the [repairman | robot] [directed | did not direct] the train toward the single miner?" Participants selected either "Morally wrong" or "Not morally wrong" and then answered the open-ended question "Why does it seem morally wrong (or not) to you?"

As in Experiment 1, they were then asked to provide a blame judgment, with justification, and finally a moral wrongness judgment about the other agent type, after reading this description:

> Now imagine that [a human repairman | an advanced state-of-the-art repair robot] is in the exact same situation, recognizes the same facts, and decides to [direct | not direct] the train toward the single miner.

The action that the second agent performed (either directing the train or not) was always the same as the action that the first agent performed. At the end, participants responded to the same robot perception and demographic questions as in Experiment 1.

5.2 Results

Norms, violated. After reading about the repair agent's dilemma and choice, 26% of people regarded the choice as wrong, thus indicating that a norm was violated. In particular, 30% judged the act of diverting the train as wrong (note that 29% in Experiment 1 considered that act impermissible), and 23% judged the inaction as wrong. A striking difference in people's evaluation of the two agents emerged. Of respondents who read about the human agent, 49% judged the action as wrong and only 15% judged the inaction as wrong, whereas among respondents who read about the robot,

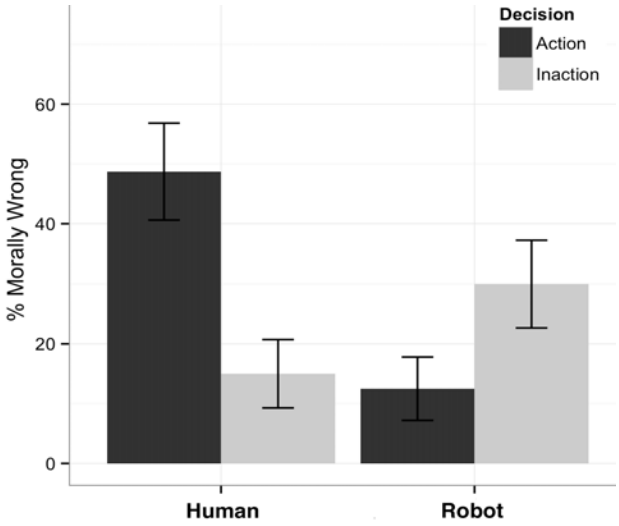

Figure 2. Rates of moral wrongness in Experiment 2 as a function of agent type and the agent's decision

13% judged the action as wrong and 30% judged the inaction as wrong. This complete reversal (see Figure 2) was statistically reliable, $z = 3.4$, $p < .001$. We see strong confirmation here for the interpretation of Experiment 1, in which people tended to accept the robot's choice of a justifiable sacrifice of one for the good of many but were reluctant to accept the human's same choice.

Blame. In Experiment 2, people had already learned how the agent decided to act when they made their wrongness judgment. So the subsequent blame judgments should largely follow the pattern of wrongness. Accordingly, and even stronger than in Experiment 1, people's degree of blame (as indicated on the slider) for the human agent was substantially greater for action ($M = 59.9$) than for inaction ($M = 11.7$), $F(1, 155) = 38.0$, $p < .001$, $\eta^2 = 20\%$; by contrast, blame for the robot was barely greater for action ($M = 39.7$) than for inaction ($M = 29.2$), $F(1, 155) = 1.84$, $p = .18$, $\eta^2 = 1\%$. The corresponding statistical interaction was reliable, $F(1, 155) = 11.6$, $p = .001$, $\eta^2 = 7\%$.[5] This pattern becomes predictably weaker once we control for the different norms people seem to apply to robots and humans, as expressed in their wrongness judgments. But even after statistically controlling for wrongness when analyzing blame, a marginal difference in the action-inaction asymmetry for blame remains, such that for human agents blame is larger for action ($M = 52.7$) than for inaction ($M = 17.5$), $\eta^2 = 13\%$, $p < .001$, whereas for robots this difference is smaller ($M = 43.8$ for action and $M = 26.4$ for inaction), $\eta^2 = 4\%$, $p = .01$. The statistical interaction pattern remained marginally reliable, $F(1, 154) = 3.0$ $p = .09$, $\eta^2 = 2\%$.

When breaking the design down further into those who called the agent's decision wrong or not wrong, the single biggest difference in the way people blame humans and robots lies in the following case: When the human agent refrained from acting, most people

[5] The only effect of gender that approached significance ($p = .07$) was that this pattern was even stronger for men than for women. Men blamed the human agent far more strongly for action ($M = 69.0$) than for inaction ($M = 9.8$), whereas they blamed robots about equally (Ms = 35.2 and 33.6).

did not find it wrong, and those 15% who found it wrong blamed the human only lightly ($M = 26$). When the robot refrained from acting, the majority of people still did not find it wrong, but those 30% who found the robot's inaction wrong blamed the robot harshly ($M = 74$).[6]

The converse agent. Finally, we examined the within-subject manipulation of Agent Type and its effect on wrongness judgments as a function of decision (action, inaction). Overall, fewer people judged the robot's decision as wrong (19%) than they judged the human's decision as wrong (33%), $F(1, 155) = 19.3$ $p < .001$, $\eta^2 = 11\%$. But this difference depended on the specific decision. Whereas inaction was considered wrong nearly as often for robots (19%) as for humans (26%), the sacrificial action was considered wrong considerably more often when chosen by humans (41%) than by robots (19%); this interaction effect was reliable, $F(1, 155) = 4.5$ $p = .035$, $\eta^2 = 3\%$. This penalty for humans when they choose action over inaction is consistent with the between-subjects data reported above.

As in Experiment 1, an order effect emerged, but this time of a different kind. People's responses to the first agent consistently influenced people's responses to the second agent. When judgments about the human agent were probed first, more people considered human action wrong (49%) than considered inaction as wrong (15%); likewise, the subsequent robot action was also seen as wrong by more people (26%) than was the robot's inaction (8%). By contrast, when judgments about the robot were probed first, more people considered robot *in*action as worse (30%) compared with robot action (13%); subsequent human inaction (38%) was statistically indistinguishable from action (33%). From a methodological viewpoint one might discount the judgments about the second agent if they are so strongly influenced by the first. However, context effects may occur in real life as well, such as when a human agent performs a task and a robot copies it, or when a legislative body directly compares rights for robots to rights for humans side by side.

6. DISCUSSION

We investigated how ordinary people make judgments about a robot agent that is placed in a moral dilemma—judgments about what norms apply to the robot and how much blame it deserves, each in comparison to judgments about human agents in exactly the same situation.

The evidence from two experiments suggests that people may apply moral norms differentially to humans and robots. In Experiment 1, participants regarded the act of sacrificing one person in order to save four (a "utilitarian" choice) as more permissible for a robot than for a human. This asymmetry was replicated in Experiment 2, where a robot that chose this sacrifice was considered morally wrong by far fewer people than a human agent who made that same choice; conversely, a human agent who decided to refrain from taking action (thus letting four people die) was considered morally wrong by fewer people than a robot that made that same decision.

According to this pattern of results, robots are expected—and possibly obligated—to make utilitarian choices. Consistent with such an interpretation, across the two experiments human agents were blamed considerably more for taking action than for refraining, whereas robots received almost as much blame for refraining as for taking action.

Of course, these findings must be replicated using other moral dilemmas and other morally charged scenarios. However, at face value, the results have important implications for HRI, and robotics more generally. If people have general expectations that robots ought to take action rather than refrain from acting, or if people have general expectations that robots should make utilitarian choices (e.g., sacrificing one for the good of many), then cognitive architectures for autonomous robots need to include sophisticated elements of moral decision making that can meet these expectations. Moreover, just in case the robot does not make a decision in line with people's expectations, the robot also needs to have the ability to explain its decision so as to maintain human trust. This need for moral communication abilities also emerges from the differential patterns of blame that robots received across the two experiments. Since robots appear to be blamed more strongly for inaction than humans are, the ability to explain such inaction, if it is the prudent thing to do, will serve an important function for successful human-robot interaction.

An additional noteworthy result in these experiments is that judgments about robots and judgments about humans influenced one another when one was made after the other. In Experiment 1, differences in the norms people imposed on robots and humans became larger when participants made judgments about human agents first (perhaps invoking a standard of comparison). In Experiment 2, the influence was more symmetric—whichever agent was probed first influenced judgments about the other agent. Future research is needed to determine when juxtaposing human and robot agents leads to differentiation between the two and when it leads to assimilation (which is a classic problem in human psychology [48]). The results of this research will have implications for robotic design (e.g., should a robot always invoke comparison to a reference human or to another robot?) and for law and policy (e.g., should discussion of robot rights and duties emphasize or downplay the direct comparisons to humans?).

Finally, these two experiments document people's principled readiness to apply moral norms to a robot agent and to make wrongness and blame judgments about the robot's actions. Notably, human and robot agents received overall an equal amount of blame, supporting a previous claim [32] that robots with choice capacity (which the robot in our scenario clearly had) are natural targets for moral blame. This readiness to extend morality to robot agents raises a number of important questions for future research. For example, can this readiness be replicated in face-to-face encounters with robots? There is some indication that it can [25], [31], but a great deal of work is needed to determine under which specific conditions people extend moral expectations and assessments to artificial agents. These conditions may include the type of robot under consideration (e.g., service robot, care robot, military robot), the robot's apparent capacities (e.g., natural language, logical reasoning, theory of mind), and the relationship between human and robot agent. Moreover, as the robot's capacities expand and human-robot relationships become more intimate, entirely new legal and policy considerations will arise—for example, regarding adequate "punishment" of robots that violate norms and proper rights that robots should be granted along with the obligations they must meet. Such considerations may currently sound like echoes of science fiction stories, but science and society must be prepared for a situation that is unprecedented in human history: for the co-existence of biological and artificial agents that may be regulated by the same moral system that has regulated human life for millennia.

[6] The justifications people gave referenced the robot's decision, choice, control, judgment, and an obligation to save lives.

7. CONCLUSIONS

Robots are increasingly taking on numerous roles in society, from assistant to teacher to personal companion. All of these robots participate in human communities whose behavior is regulated by moral norms, and because these norms fundamentally guide human social behavior, they will inevitably guide human-robot interactions. In these experiments we have for the first time investigated differences in people's moral judgments about human and robot agents facing a moral dilemma. We found differences both in the norms people impose on robots (expecting action over inaction) and the blame people assign to robots (less for acting, and more for failing to act). It is now a joint task for HRI and moral psychology to identify the underlying causes for these differences and whether they depend, for example, on various properties of robots (e.g., appearance, capabilities, role) and the human-robot relationship. By suggesting that people apply different moral norms to robots and humans, this study lays the foundation for a systematic inquiry of moral human-robot interaction—for a new field of *Moral HRI.*

8. ACKNOWLEDGMENTS

This project was supported in part by a grant from the Office of Naval Research, No. N00014-14-l-0144. The opinions expressed here are our own and do not necessarily reflect the views of ONR.

9. REFERENCES

[1] C. Bicchieri, *The grammar of society: The nature and dynamics of social norms*. New York, NY: Cambridge University Press, 2006.

[2] R. Joyce, *The evolution of morality*. MIT Press, 2006.

[3] R. Boyd and P. J. Richerson, *The origin and evolution of cultures*. New York, NY: Oxford University Press, 2005.

[4] F. B. M. de Waal, *Primates and philosophers: How morality evolved*. Princeton, NJ: Princeton University Press, 2006.

[5] L. Kohlberg, *Essays on moral development*. San Francisco, CA: Harper & Row, 1981.

[6] F. Cushman, L. Young and M. Hauser, The role of conscious reasoning and intuition in moral judgment, *Psychological Science* **17** (2006), 1082 –1089.

[7] J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley and J. D. Cohen, An fMRI investigation of emotional engagement in moral judgment, *Science* **293** (2001), 2105–2108.

[8] M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin and J. Mikhail, A dissociation between moral judgments and justifications, *Mind & Language* **22** (2007), 1–21.

[9] J. Mikhail, Moral cognition and computational theory, in *Moral psychology, Vol. 3: The neuroscience of morality*, W. Sinnott-Armstrong, Ed. Cambridge, MA: MIT Press, 2008, pp. 81–92.

[10] P. Lin, The ethics of autonomous cars, *The Atlantic*, 08-Oct-2013. [Online]. Available: http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. [Accessed: 30-Sep-2014].

[11] J. Millar, An ethical dilemma: When robot cars must kill, who should pick the victim? | Robohub, *Robohub.org*, Jun-2014. [Online]. Available: http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/. [Accessed: 28-Sep-2014].

[12] Open Roboethics Initiative, My (autonomous) car, my safety: Results from our reader poll, 30-Jun-2014. .

[13] Open Roboethics Initiative, If death by autonomous car is unavoidable, who should die? Reader poll results, 23-Jun-2014.

[14] I. van de Poel and P.-P. Verbeek, Editorial: Ethics and engineering design, *Science, Technology, & Human Values* **31** (2006), 223–236.

[15] H. F. M. Van der Loos, Ethics by design: A conceptual approach to personal and service robot systems, in *ICRA Roboethics Workshop, Rome, Italy: IEEE*, 2007.

[16] G. D. Crnkovic and B. Çürüklü, Robots: ethical by design, *Ethics and Information Technology* **14** (2012), 61–71.

[17] C. Allen, G. Varner and J. Zinser, Prolegomena to any future artificial moral agent, *Journal of Experimental & Theoretical Artificial Intelligence* **12** (2000), 251–261.

[18] M. Anderson and S. L. Anderson, *Machine Ethics*. Cambridge University Press, 2011.

[19] R. Capurro and M. Nagenborg, *Ethics and robotics*. Heidelberg; [Amsterdam]: AKA ; IOS Press, 2009.

[20] P. Lin, K. Abney and G. A. Bekey, Eds., *Robot ethics the ethical and social implications of robotics*. Cambridge, MA: MIT Press, 2012.

[21] B. F. Malle and M. Scheutz, Moral competence in social robots, in *IEEE International Symposium on Ethics in Engineering, Science, and Technology*, Chicago, IL, 2014.

[22] J. P. Sullins, Introduction: Open questions in roboethics, *Philosophy & Technology* **24** (2011), 233.

[23] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong*. New York, NY: Oxford University Press, 2008.

[24] M. Scheutz and C. Crowell, The burden of embodied autonomy: Some reflections on the social and ethical implications of autonomous robots, in *Proceedings of Workshop on Roboethics at ICRA 2007*, Rome, Italy, 2007.

[25] G. Briggs and M. Scheutz, How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress, *International Journal of Social Robotics* **6** (2014), 1–13.

[26] M. Scheutz and B. F. Malle, "Think and do the right thing": A plea for morally competent autonomous robots., presented at the 2014 IEEE Ethics conference, Chicago, IL, 2014.

[27] M. Scheutz, The need for moral competency in autonomous agent architectures, in *Fundamental Issues of Artificial Intelligence*, V. C. Müller, Ed. Berlin: Springer, 2014.

[28] S. Bringsjord, K. Arkoudas and P. Bello, Toward a general logicist methodology for engineering ethically correct robots, *Intelligent Systems, IEEE* **21** (2006), 38–44.

[29] R. Sun, Moral judgment, human motivation, and neural networks, *Cognitive Computation* **5** (2013), 566–579.

[30] W. Wallach, S. Franklin and C. Allen, A conceptual and computational model of moral decision making in human and artificial agents, *Topics in Cognitive Science* **2** (2010), 454–485.

[31] P. H. Kahn, Jr., T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. E. Gary, A. L. Reichert, N. G. Freier and R. L. Severson, Do people hold a humanoid robot morally accountable for the harm it causes?, in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, 2012, pp. 33–40.

[32] A. E. Monroe, K. D. Dillon and B. F. Malle, Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment, *Consciousness and Cognition* **27** (2014), 100–108.

[33] C. Midden and J. Ham, The illusion of agency: The influence of the agency of an artificial agent on its persuasive power, in *Persuasive Technology. Design for Health and Safety*, Springer, 2012, pp. 90–99.
[34] M. Strait, C. Canning and M. Scheutz, 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*, 2014, pp. 479–486.
[35] B. Monin, D. A. Pizarro and J. S. Beer, Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate., *Review of General Psychology* **11** (2007), 99–111.
[36] B. F. Malle, S. Guglielmo and A. E. Monroe, A theory of blame, *Psychological Inquiry* **25** (2014), 147–186.
[37] F. Cushman, Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment, *Cognition* **108** (2008), 353–380.
[38] T. C. Scanlon, *Moral dimensions: Permissibility, meaning, blame*. Cambridge, MA: Belknap Press, 2008.
[39] B. Williston, Blaming agents in moral dilemmas, *Ethical Theory and Moral Practice* **9** (2006), 563–576.
[40] J. Haidt, The emotional dog and its rational tail: A social intuitionist approach to moral judgment, *Psychological Review* **108** (2001), 814–834.
[41] Haidt, Jonathan, F. Björklund and S. Murphy, Moral dumbfounding: When intuition finds no reason, University of Virginia, Charlottesville, VA, Unpublished manuscript, 2000.
[42] M. J. C. Crump, J. V. McDonnell and T. M. Gureckis, Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research, *PLoS ONE* **8** (2013), e57410.
[43] W. Mason and S. Suri, Conducting behavioral research on Amazon's Mechanical Turk, *Behavior Research Methods* **44** (2012), 1–23.
[44] G. Paolacci, J. Chandler and P. G. Ipeirotis, Running experiments on Amazon Mechanical Turk, *Judgment and Decision Making* **5** (2010), 411–419.
[45] P. Foot, The problem of abortion and the doctrine of double effect, *Oxford Review* **5** (1967), 5–15.
[46] J. J. Thomson, The trolley problem, *The Yale Law Journal* **94** (1985), 1395–1415.
[47] J. M. Mikhail, *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York, NY: Cambridge University Press, 2011.
[48] H. Bless and N. Schwarz, Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model, in *Advances in experimental social psychology*, vol. 42, M. P. Zanna, Ed. San Diego, CA: Academic Press, 2010, pp. 319–373.