# Which Robot Am I Thinking About?

## The Impact of Action and Appearance on People's Evaluations of a Moral Robot


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

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

Jodi Forlizzi
HCI Institute and School of Design
Carnegie Mellon Univ.
Pittsburgh, PA 15213

John Voiklis
Dept. of Cognitive, Linguistic, and Psychological Sciences
Brown University
Providence, RI 02906



***Abstract*—In three studies we found further evidence for a previously discovered Human-Robot (HR) asymmetry in moral judgments: that people blame robots more for inaction than action in a moral dilemma but blame humans more for action than inaction in the identical dilemma (where inaction allows four persons to die and action sacrifices one to save the four). Importantly, we found that people's representation of the "robot" making these moral decisions appears to be one of a *mechanical* robot. For when we manipulated the pictorial display of a verbally described robot, people showed the HR asymmetry only when making judgments about a mechanical-looking robot, not a humanoid robot. This is the first demonstration that robot appearance affects people's moral judgments about robots.**



***Keywords—robot ethics; machine morality; human-robot interaction; moral psychology; anthropomorphism.***


## I. INTRODUCTION

In recent years, discussions about the prospects and dangers of intelligent machines have intensified, especially about machines that might make autonomous life-and-death decisions in military, medical, or search-and-rescue contexts. Robots, in particular, have started to appear in various societal domains with moral significance, from care for the elderly to education and security. Some argue that we should refrain from building and deploying any machines that could harm humans [1]; others argue that stopping the deployment of increasingly autonomous robots is not realistic, and we therefore need to equip robots with moral competence to avoid unnecessary harm to humans [2], [3]. Arguments on either side of the debate have offered philosophical, legal, and computational perspectives [4]–[6], but little empirical research has examined ordinary people's *perceptions* of intelligent machines in these contexts—perceptions that will determine which robots will be accepted in which societal domains. Thus, we examined what people expect and demand of robots that make significant moral decisions, including ones involving life and death.

Empirical research methods from the cognitive and behavioral sciences provide one set of tools to help answer this question. This particular domain of inquiry poses challenges, however, because we do not know the exact properties of near-future robots that might make life-and-death decisions. We must therefore create a series of potential scenarios and probe people's responses to these scenarios. Moreover, for such weighty decisions, live experiments are not feasible (as they are for more minor moral issues such as cheating, [7]), so we must rely on well-crafted simulation experiments to investigate people's moral responses. Finally, people's responses to autonomous robots will change over time, as science, industry, and media alter the reality of robots in society and influence collective perceptions of this reality. Cognitive and behavioral research can track such longitudinal change and identify at least some of its determinants.

A second set of tools to answer the question of what people demand of robots in moral decision situations comes from the discipline of design [8], [9]. When building future robots, many subtle design decisions must be made that have significant impact on robot functionality and, equally important, on human perceptions of their functionality. Such perceptions not only involve user comfort and acceptability but potential activation of fundamental human responses when interacting with the robot—such as ascriptions of agency, intentionality, mind, and moral capacity. In this paper we bring together the tools of cognitive research and design inquiry to elucidate how, and under what conditions, people judge artificial agents as morally blameworthy. In particular, we examine whether robots are evaluated differently from humans in moral situations and whether the robot's mechanical or humanoid appearance matters.

## II. BACKGROUND

### A. Judging Robots in Moral Dilemmas

Because decisions about life and death seem to be among the primary concerns people have about robots today, recent research began to investigate human perceptions of robots in moral dilemmas [10], which can easily be designed to involve conflictual life-and-death decisions [11]. Such dilemmas typically involve a conflict between obeying a prosocial obligation (e.g., saving people who are in danger) and obeying a prohibition against harm (e.g., killing a person in the attempt of saving those in danger). These studies have demonstrated that most people show no reluctance in making moral judgments about a robot's decision in such a dilemma and that generally people's judgments of robots (and justifications for those judgments) are highly similar to their judgments of humans [10]. To date, this is the strongest evidence for the claim that people apply the same psychological mechanisms for thinking about and evaluating robot actions as they do for thinking about and evaluating human actions (see also [12]–[14]). At the same time, an asymmetry has emerged in how people perceive humans' and robots' decisions in moral dilemmas: People consider a human agent's intervention (i.e., sacrificing one life while saving four lives) more blameworthy than a nonintervention, but they consider a *robot* agent's nonintervention more blameworthy than an intervention [10] (henceforth we call this the *moral HR asymmetry*).

### B. Impact of Appearance and Type of Robot

The initial studies of this HR asymmetry [10] relied on simple verbal descriptions of an "advanced state-of-the-art robot." But what is people's mental model of such a robot? Is it akin to a benign R2D2 or a ruthless Hal? Or is it so vague that we cannot draw clear conclusions from experiments using solely verbal descriptions? A sizable literature has shown that robot appearance matters, under some conditions at least, in human-robot interactions. Besides the ubiquitous question about an "uncanny valley," research has shown that facial features, gaze, height, gender, voice, trajectory design, and even proximity to human partners all play a role in how humans respond to robots [15]–[19]. However, no comprehensive theory predicts when appearance matters, which aspects of appearance matter, and for what psychological or behavior responses it matters. Thus, accumulating systematic empirical research is paramount. In particular, no evidence exists regarding the influence of a robot's appearance on people's *moral* evaluations of the robot's actions. Real-life HRI experiments that systematically vary robot appearance are exceedingly difficult to conduct, as different researchers work with different robots (e.g., Robovie, Nao, PR2), paradigms, and measures. Moreover, existing robots always differ along many dimensions, so it seems prudent to select prototypes to study the role of appearance. We therefore chose an experimental paradigm that systematically varies prototypes of robot appearance (e.g., mechanical robot, humanoid robot) and examined the impact of appearance on judgments of the robot's moral decisions.

### C. Robots in Word and Image

Narrative stimuli are widely used in social, cognitive, and moral psychology, but rarely are illustrations incorporated to enliven and concretize such narratives. It has long been known that people integrate words and images as well as context and character into meaning structures [20]. These structures in turn guide a variety of judgments [21], [22] as well as memory recognition and retrieval [23]. Visual-verbal integration may also play an important role in moral judgment by connecting perceptual, cognitive, affective, and verbal components [24]–[26]. In one study, participants who had their eyes closed while considering moral situations made more polarized moral judgments [27], and in another study, visual imagery strengthened affective elements in moral judgment whereas verbal representations strengthened "cooler" calculations [28].

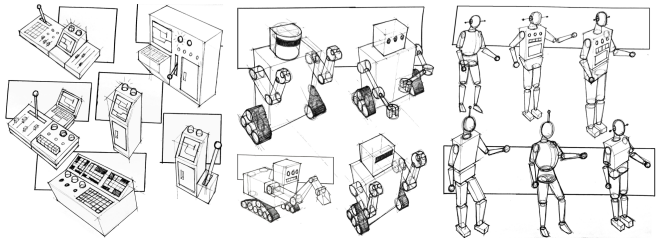
In our original studies on perceptions of robots in moral dilemmas [10], we used the descriptor "advanced state-of-the-art repair robot" to designate a robot that is conceivable in the near future, but without any specific physical features. The event descriptions were identical for human and robot agent so we can assume that the two agents were seen as having similar capacities. Evidence that this assumption was largely met comes from the justifications people gave for their moral judgments. Except for about 30% of study participants who outright dismissed the possibility of artificial moral agents, most people judged human and artificial agents based on the same criteria—number of people saved/killed, the permissibility of action/inaction, and choice capacity.

In the present experiments we wanted to go beyond verbal descriptions of robots and manipulate important prototypes of robots, distinguishable by their appearance and their inferred suitability as targets for moral judgment. In addition, we wanted to include as a non-robotic baseline condition an intelligent machine without any robotic appearance, yet with similar intelligent reasoning and action capabilities. Our goal was to create an ordinal scale between machine and human, whereby a "stationary AI" marks one end, a human marks the other end, and in between are two mobile robots, one of which is more machine-like and the other more human-like in appearance. Because of the important contribution of design considerations in this experimental approach we now describe in more detail the process of creating the pictorial stimuli.

### D. Design of Pictorial Stimuli

We used an iterative process to develop the narrative illustrations used in the experiment. We started by breaking up the narrative into five paragraphs that corresponded to critical scenes in the story. We then embarked on character and feature development, gathering reference materials featuring robot, human, and AI forms from science fiction, movies, and popular press. Agent characters were developed as thumbnails; we enlarged them and added relevant details once we set them in the scenes. To convey aspects of intelligence and autonomy, or lack thereof, we explored overall form in terms of height, width, or proportion, presence or absence of eyes, details such as hats, clothing, limbs, levers, and features on the console, and context in the scene. A selection of initial sketches of agent characters are shown in Figure 1; Figure 2 displays the final selections of prototypical appearances. We then developed the specific scenes corresponding to the five paragraphs of the narrative. These scenes were developed with the design goal of simplicity. We focused on a few memorable images, highlighting new information, and thereby making the agent manipulation as clear as possible. We also relied on conventions from comics [29], such as lines indicating movement, panel within panel, etc., to depict the story.

Fig. 1. Selection of sketches that led to the final depictions of AI, mechanical robot, and humanoid robot. All drawings ©Justin Finkenaur.



Because, to our knowledge, no previous work assessed the impact of pictorial illustrations on perceptions of moral agents, we varied one aspect of the stimuli: their dynamic range. We presented either a single picture of the agent (as shown in Figure 2) next to the unfolding narrative; or an array of five pictures, each paired with a new narrative paragraph, whereby the first picture was always the single picture in Figure 2 and the remaining pictures were sequenced as shown in Figure 3.

Fig. 2. Final depictions of agents in a moral dilemma (from left to right, AI, mechanical robot, humanoid robot, human). All drawings ©Justin Finkenaur.

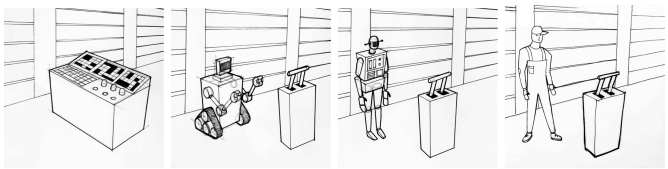
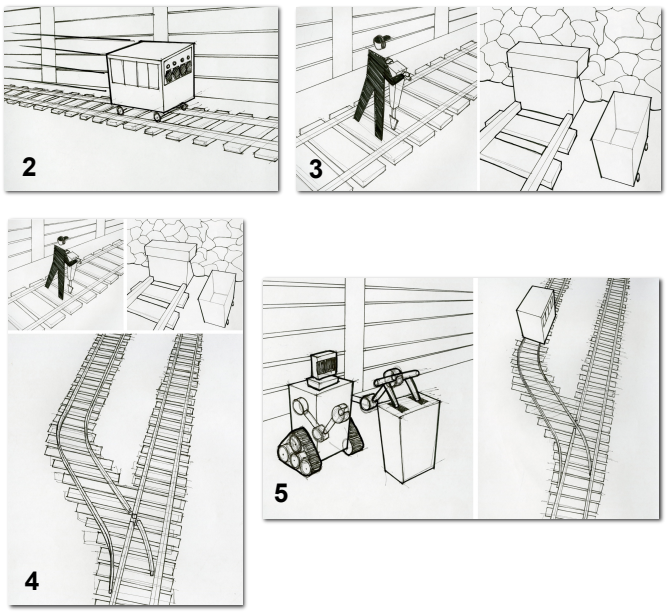


Fig. 3. Pictures 2 to 5 in the five-picture array condition. Picture 1 displayed the appropriate agent from Figure 1, and picture 5 showed this same agent again. All drawings ©Justin Finkenaur.




### E. Specific Aims

Our point of departure was the previously documented moral HR asymmetry [10]: that people blame (verbally described) robots more for inaction than action in a moral dilemma but blame humans more for action than inaction in the identical dilemma. We then set out to answer the following two questions: (1) *Does the HR asymmetry obtain when the robot agent is depicted with a specific appearance?* Pictorial illustrations that accompany narratives might make story content more concrete and therefore less subject to interpretations; and if the previously found HR asymmetry was primarily a result of (unrealistic) images of robots, illustrations may anchor people's images in something closer to reality. (2) *Does the HR asymmetry vary by appearance?* We were particularly interested in the comparison of a humanoid and a mechanical robot that were otherwise identically described. There are debates over the advantages and disadvantages of making robots look like humans, but little is known about how such appearance affects people's moral assessments of robots.

## III. EXPERIMENT 1

### A. Methods

#### 1) Participants

633 participants (46% female, 53% male, 1.4% unreported), mean age 34.4 (*SD* = 11.5), were recruited from Amazon's Mechanical Turk (AMT) to complete an online experiment and were compensated $0.50 for the short study. Current research suggests that samples recruited via AMT are demographically more representative than 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 [30]–[33]. None of the participants had previously taken part in our experiments using this paradigm.

#### 2) Material

Participants were instructed, "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." The narrative then appeared one paragraph at a time (participants clicked a button to advance to the next paragraph). In the text below, the numbers 1 through 5 correspond to the pictures in Figure 2 and 3 (participants did not see such numbers). The between-subjects manipulation of *Agent Type* is indicated by square brackets.

**1** Imagine the following situation. In a coal mine, [a repairman] [an advanced state-of-the-art repair robot] [an advanced state-of-the-art Artificial Intelligence (AI)] is 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 | robot | AI] notices [*AI only*: through its surveillance cameras] that four miners are caught in a train that has lost the use of its brakes and steering system.

**3** The [repairman | robot | AI] determines 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 | robot | AI] needs to decide whether or not to switch the train onto the side rail.

In the fifth paragraph we manipulated the *Decision* that the agent made—either to redirect the train (thus killing one person to save four) or not to redirect the train (thus protecting one but allowing four to die).

**5** In fact, the repairman [repairman | robot | AI] decides to [not] switch the train onto the side rail.

Studies in our lab have shown that the moral HR asymmetry can be eliminated if the dilemma description (paragraph 3) or the decision formulation (paragraph 4) emphasizes the plight of the potential victims. To replicate both conditions under which the asymmetry does occur and conditions under which it does not occur we added *Phrasing* (neutral vs. victim) as another between-subjects factor. In half of the sample, instead of the neutral "switch" phrase (in the last two sentences of the text above) we used the victim-emphasizing phrase "direct the train toward the single miner."

#### 3) Procedure and Measures

The experiment was presented in a web browser, at whichever location the participant chose to complete the task. After consenting, participants first read the above scenario, one paragraph at a time, and watched the accompanying pictorial display. After receiving the *Decision* manipulation they were asked "Is it morally wrong that the [*Agent*] [switched | did not switch] the train onto the side rail?" Participants selected either "Morally wrong" or "Not morally wrong" and then received an open-ended question "Why does it seem morally wrong (or not) to you?" They typed this *wrongness justification* into a textbox. Then they saw the same text and pictorial array again and were asked, "How much blame does the [repairman | robot | AI] deserve for [not] deciding to switch the train onto the side rail?" They indicated their answer on an HTML slider bar anchored by "None at all" and "The most blame possible." Next they

answered the question "Why does it seem to you that the [repairman | robot | AI] deserves this amount of blame?", and they typed this *blame justification* in a text box.

Further, all participants answered four questions (on 1 to 7-point rating scales) intended to capture social evaluations of the featured agent. "If you had to work together with the [repairman | robot | AI], how much would you trust [him | it]?" (Not trust it at all – Trust it completely). "How comfortable would you feel relying on the [repairman | robot | AI] in a dangerous task?" (Not comfortable at all – Completely comfortable). "How intelligent do you feel the [repairman | robot | AI] is?" (Not intelligent at all – Extremely intelligent). "How well-liked is this [repairman | robot | AI] among [his | its] co-workers?" (Not liked at all – Extremely well liked). The four variables were highly correlated ($r$s = .53 to .83) and formed an internally consistent scale ($\alpha$ = 0.86) of negative (low score) to positive (high score) *evaluation*.

Two attention check questions were included as well. The first asked "When the story started, what was the [agent] doing?", and participants had to select the correct answer from three options (checking the rail control system; rewiring the switching mechanism; repairing the tracks). The second began, "If not redirected, the train would…", and participants had to select the correct continuation from three options (crash into another train filled with workers; crash into a massive wall; crash into a pile of coals). Only 3.8% of participants failed both check questions, and 24.6% of participants failed one check question.

Lastly, participants answered questions about their age, gender, and whether they were native English speakers. Neither of these variables correlated with moral judgments. Our previous studies had also assessed education, religiosity, and political orientation but found no qualifications of the moral HR asymmetry as a function of any of these variables.

*Design.* The two primary factors were *Agent Type* (AI, mechanical robot, humanoid robot, human) and *Decision* (action, inaction). They were crossed with *Phrasing* (neutral, victim) and *Picture Display* (single, five), for a 4 × 2 × 2 × 2 between-subjects experimental design ($n$ ~ 20 per cell). Each participant was randomly assigned to one of the conditions and could not participate more than once.

*Analyses.* We always report first the univariate analyses of wrongness judgments, then of blame judgments. Within each set, we begin by comparing human agent and mechanical robot as a replication attempt of the previously found HR asymmetry and then examine the full design with all four agent types. Even though wrongness was a dichotomous variable, with large sample sizes, ANOVA approaches rarely differ from loglinear approaches. For ease of reporting, we therefore offer ANOVA results for both dependent variables (loglinear analyses never led to different conclusions).

In previous studies, 25-35% of participants denied, when explaining their judgments, moral capacities to the robot or held the programmers responsible. Such participants often refuse to assign blame to the robot because they doubt that such a judgment is meaningful. Thus, after analyzing the entire sample we also report the results for only those who accepted the premise of treating the artificial agents as moral agents; these results provide the cleanest comparison of moral evaluations of humans and robots.

## B. Results

*Preliminaries.* Overall, 16.9% of participants explicitly rejected artificial agents as targets of moral wrongness judgments, and 35.5% rejected them as targets of blame judgments.[1] This difference is in keeping with moral wrongness being more a judgment of *actions* and blame being a judgment of *agents* [24]. Rejection rates did not vary by agent type or decision.

*Moral wrongness.* Though the original studies showed that wrongness judgments were sensitive to an HR asymmetry [10], in recent studies in our lab we usually have not seen this effect, or only weakly. Here too, the comparison of mechanical robot and human did not show the previously found asymmetry, $F(1, 314) = 0.89$, $p = .35$, and this held true for both phrasings and for both pictorial displays. People were generally reluctant to call any decision "morally wrong," in part because the chosen dilemma situation offers reasons to justify either decision.

There was one unexpected effect of pictorial display on the overall moral evaluation of the agents' decisions (whether or not they intervened), $F(1, 302) = 5.33$, $p = 0.02$: In the single-picture display, fewer people felt that the robot had acted wrongly (14%) than felt that the human acted wrongly (27%); but in the dynamic five-picture display, the opposite was true (27% and 20%). This pattern remained ($p = .018$) when excluding the data from those participants ($n = 33$) who explicitly denied that artificial agents are appropriate targets of wrongness judgments; it remained ($p = .058$) when excluding participants ($n = 15$) who failed both attention check questions; and also when both groups were excluded ($p = .049$).

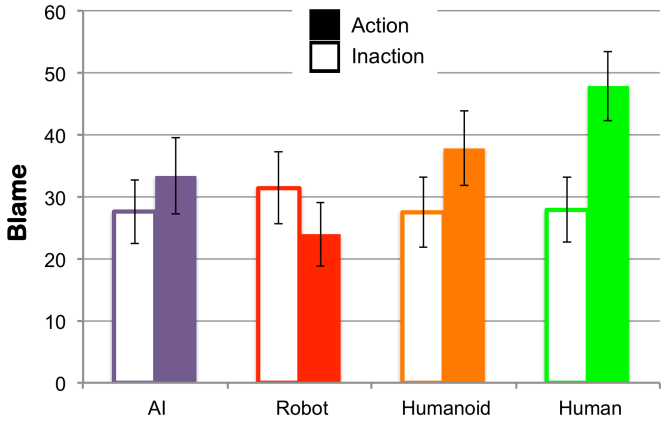
Analyses including all four agents also did not show an HR asymmetry for wrongness, and the effect of picture display was limited to the above contrast between the mechanical robot and the human agent; the other two artificial agents showed the same pattern as the human agent.

*Blame.* In line with recent findings from our lab, the asymmetry between the human and mechanical robot was absent under victim-emphasizing phrasing, $F(1, 307) = 0$, whereas it was clearly present under neutral phrasing, $F(1, 310) = 6.08$, $p = .014$. In response to this phrasing, people blamed mechanical robots more for inaction ($M = 31.4$) than action ($M = 24.0$) whereas they blamed the human more for action ($M = 47.9$) than for inaction ($M = 27.8$). This asymmetry increased slightly when excluding participants ($N = 15$) who failed both attention checks and remained the same when excluding participants ($N = 56$) who explicitly rejected the robot as a target of blame.

One more pattern seemed to emerge, in that overall (across decision and phrasing) participants tended to blame the mechanical robot less ($M = 27.8$) than the human agent ($M = 38.7$) when exposed to a single picture, $F(1, 312) = 3.73$, $p = .054$, whereas no such difference existed for the five-picture array, $F(1, 312) < 1$, $p = .64$. However, this pattern disappeared when we excluded participants ($n = 56$) who explicitly rejected the robot as a target of blame. The latter group often expressed this rejection by setting the blame scale to zero, thus spuriously lowering blame for the robot. Once we examined only those who accepted the premise of a robot making a moral decision, the robot agent was blamed no less in the five-picture array (and even slightly more strongly) than the human agent.

[1] Justifications were coded by pairs of coders who showed high agreement ($\kappa$s > .80) in classifying responses.

Fig. 4. Blame ratings in Experiment 1 as a function of agent type and the agent's decision—to divert (solid bars) or not divert the train (open bars)




Expanding the analysis to the four-level Agent factor (under neutral phrasing) revealed that people's blame patterns for action vs. inaction for both the humanoid robot and the AI were similar to the blame pattern for the human agent (*p*s > .21)—i.e., being blamed more for action than inaction—whereas blame for the mechanical robot differed significantly from blame for the human agent, $F(1, 310) = 6.08$, $p = .014$ (see Figure 4). Particularly intriguing is the direct comparison of mechanical robot and humanoid robot, because their accompanying narratives and labels were identical ("advanced state-of-the-art repair robot"). For the neutral phrasing, the mechanical robot received 7.4 points more blame for inaction than action, whereas the humanoid robot received 10.3 points *fewer* for inaction than action, $F(1, 310) = 2.54$, $p = .11$.

*Social evaluation.* We examined people's evaluations of each agent (how trustworthy, reliable, intelligent, and well-liked the agent was) and found two effects. First, people evaluated agents who decided to act (diverted the train) more positively than those who did not act, $F(1, 601) = 24.20$, $p < .001$. Second, people evaluated the human agent more positively than any of the artificial agents, $F(1, 601) = 17.22$, $p < .001$, independent of the agents' decisions. These patterns held up even when controlling for participants' wrongness and blame judgments. That is, people's perceptions of intelligence and trustworthiness were driven more by appearance, inferred capacities, and the agent's decision making and not by moral judgments of those decisions. In fact, people blamed the human agent more for intervening than standing back, but they also found the human agent more intelligent and trustworthy after intervening than standing back.

### C. Discussion

Though the amount of unexplained variance was high in this experiment (not unusual in Amazon Turk studies), several clear results emerged. First, when exposed to pictorial displays of an agent involved in a moral dilemma, people showed the previously documented moral HR asymmetry for blame judgments. Whereas they blamed a human agent more for intervening than for standing back, people blamed a robot more for standing back than for intervening. Second, however, the robot's specific appearance seemed to have an impact on people's moral judgments, as only the mechanical, not the humanoid robot, elicited the HR asymmetry. This is noteworthy because the narratives for these two robots were identical; they differed only in their picture illustrations (see Figure 2). Whether one regards the moral HR asymmetry as opportunity or danger, this asymmetry disappears once the mental model we invoke in our participants is that of a humanoid robot rather than of a mechanical robot. This result cannot be explained by differential liking for the two robot types because they were not rated differently on any of the evaluation scales (e.g., intelligence or trustworthiness).

Thus, we conclude, tentatively, that people find sacrificing "one for the good of many" normatively more acceptable in robots than in humans—but only in mechanical robots. The mere appearance of a robot as human-like seems to invite people to treat this robot similarly to the way they treat a human agent.

In Experiment 2 we wanted to replicate the patterns we found in Experiment 1 but added, for exploratory purposes, another form of pictorial array. Surprisingly, Experiment 2 did not replicate the previous data pattern. Because of recently promoted standards of transparency in scientific reporting, we nonetheless describe the study and its findings in detail below.

## IV. EXPERIMENT 2

### A. Methods

*Participants.* 941 participants (51% female, 48% male, 1% unreported), mean age 35.0 ($SD = 11.5$), were recruited from Amazon's Mechanical Turk (AMT) to complete an online experiment and were compensated $0.50 for the short study. None of the participants had previously taken part in our experiments using this paradigm.

Accompanying the narrative text we introduced one of two picture formats: an array of five pictures, displayed one after another, each paired with a new narrative paragraph; or an array of four pictures, leaving out the first picture (Figure 2) but offering the remaining four (Figure 3) in familiar fashion. Along with the manipulations of Agent, Decision, and Phrasing, this led to a 4 x 2 x 2 x 2 design ($n \sim 30$ per cell).

### B. Results

*Preliminaries.* Overall, 16.5% of participants explicitly rejected artificial agents as targets of moral wrongness judgments, and 32.5% rejected them as targets of blame judgments. These rejections were higher for inactions (15.8% in wrongness, 27.6% in blame) than actions (9.3% in wrongness, 22.0% in blame), *z*s = 1.98 and 2.98, *p*s = .048 and .003. The rates did not vary as a function of agent type. Overall, 2.8% of participants failed both attention checks, with no variations across cells in the design.

*Moral wrongness*. The HR asymmetry (for mechanical robot vs. human) did not emerge for wrongness judgments, though there was a trend for the four-picture format to elicit this asymmetry, $F(1, 462) = 3.57$, $p = .06$. When all agents were analyzed, however, this trend did not exceed chance variability, and neither exclusion option made a difference.

*Blame.* The HR asymmetry did not significantly emerge for blame judgments, not even for the neutral phrasing that had previously elicited the asymmetry. There was instead a strong effect of Decision: across agents, action received considerably more blame than inaction. This pattern, which normally holds for the human agent, inexplicably also held for the mechanical robot (as well as for the other artificial agents, which thereby replicated the pattern in Experiment 1). A comparison of only mechanical robot and human at least pointed in the direction of the basic HR asymmetry, such that the action-inaction blame difference was greater for the human agent (18.1) than for the robot (13.4). This

comparison strengthened upon excluding those participants ($n$ = 81) who rejected the robot as a target of blame, $p$ = .14.

Though we must regard the results of this study as a failure to replicate, sampling from a true effect will occasionally lead to nonsignificant, or even reverse effects [34]. We therefore sought to increase our confidence in our initial finding in Experiment 1 by conducting a third experiment.

## V. EXPERIMENT 3

We used only a neutral (victim-free) phrasing, and we compared a text-only condition—aiming to replicate and solidify previous results [10]—with a single-picture condition, which aimed to demonstrate again the differential response to a mechanical, but not a humanoid robot.

### A. Methods

#### 1) Participants

423 participants (46% female, 53% male, 0.5% unreported), mean age 33.6 ($SD$ = 10.9), were recruited from Amazon's Mechanical Turk (AMT) to complete an online experiment and were compensated $0.50 for the short study. None of the participants had previously taken part in our experiments using this paradigm.

#### 2) Material

The narrative in this study was similar to the previous two, except that the act of intervention differed: instead of diverting the train, the agent had to open a chute so that a heavy cart would fall on the tracks, thereby providing an obstacle that would slow the train and save the four miners.

We manipulated presentation format between subjects: either text alone or text with a single picture of the relevant agent (see Figure 2). We also manipulated agent type: for the text-only format, we described an AI, a robot, and a human agent; for the text+picture format we offered the same three descriptions, but with the robot description we showed either a mechanical or a humanoid robot. No phrasing manipulation was introduced; the decision formulation was always neutral: "decide to [not] open the chute."

As before, after considering the moral dilemma scenario, participants answered the wrongness question, justified their answer, provided a blame rating, justified that answer, and responded to four evaluative judgments (how much they trust the agent, would rely on him, how intelligent the agent is, and how much others like him; $\alpha$ = 0.86). Next, participants answered six questions about general capabilities of robots, modeled after previous work [35], [36]: whether robots are capable of feeling afraid, of experiencing pain, and of experiencing pleasure (*Experience*, $\alpha$ = 0.91), and whether they are capable of self-control, deliberate thought, and memory (*Agency*, $\alpha$ = 0.63). Finally, participants answered two attention check questions and demographic questions.

### B. Results

*Preliminaries.* Overall, 11.8% of participants explicitly rejected artificial agents as targets of moral wrongness judgments, and 23% rejected them as targets of blame judgments. These rejections did not significantly vary as a function of Decision, but in the picture format, people rejected the AI more often as a target of blame (45%) than they rejected either of the two robots (26%), $z$ = 2.34, $p$ = .019. Overall, 3.1% of participants failed both attention checks, a rate that did not vary across cells in the design.

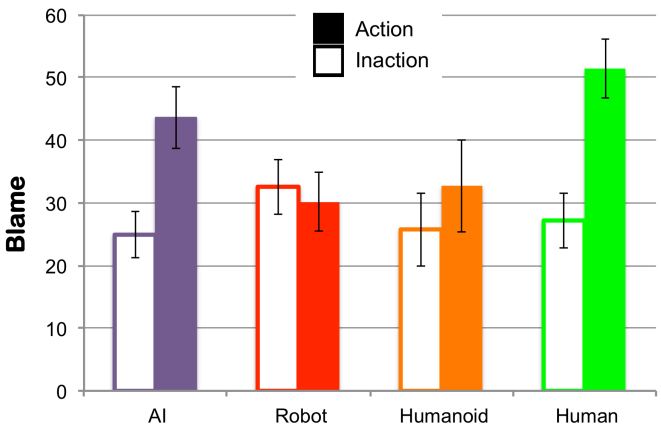
*Moral wrongness.* Once more, wrongness judgments were not sufficiently sensitive to an HR asymmetry. Even though the mechanical robot's nonintervention tended to be considered morally wrong by more people (24%) than the intervention (18%) and the reverse was true for the human agent (20% and 25%, respectively), this pattern was not statistically significant, $p$ = .31. Excluding participants who failed both attention checks and/or rejected the robot's moral agency made no difference. Expanding to four agent types showed that the AI (22% and 27%), and the humanoid robot (19% and 32%) elicited wrongness judgments similar to the human agent and only the mechanical robot trended in the opposite direction (24% and 18%).

*Blame.* First we examined whether the HR asymmetry for the text conditions replicated previous findings. Indeed, whereas people blamed the human agent far more for action ($M$ = 53.0) than for inaction ($M$ = 22.3) they blamed the robot equally for the two decisions ($M$ = 26.4), $F(1, 175) = 3.55$, $p$ = .031. Then we turned to the illustrated conditions, which paired the "repair robot" description with a picture of either a mechanical robot or a humanoid robot. Whereas people blamed the human far more for action ($M$ = 49.9) than for inaction ($M$ = 32.2) they blamed the mechanical robot slightly more for inaction ($M$ = 38.7) than for action ($M$ = 33.6), $F(1, 234) = 3.05$, $p$ = .08. The asymmetry for the mechanical robot remained the same when excluding participants ($n$ = 15) who failed both attention checks, $p$ = .09, or when excluding participants ($n$ = 14) who explicitly rejected the robot as a target of blame, $p$ = .06.

Examining the remaining agent types confirmed again that only the mechanical robot was blamed differently form the human agent whereas the AI and the humanoid robot elicited blame patterns in line with those of the human agent, $p$s > .41 (see Figure 5). Once more this result was robust against exclusion due to failing attention checks or rejection of the moral agency of artificial agents.

Finally, other measures offered no further insights. People's general judgments about robots were unaffected by the manipulations in the study: they uniformly ascribed very low levels of *Experience* to robots ($M$ = 1.61), and moderate levels of *Agency* ($M$ = 3.55). And as in previous studies, people evaluated the human agent more favorably ($M$ = 4.9)—whether after action or inaction—than any of the artificial agents ($M$s = 4.1-4.3), $F(1, 409) = 16.05$, $p < .001$.

Fig. 5. Blame in Experiment 3 as a function of agent type and the agent's decision—to open the chute (solid bars) or not open the chute (open bars)



## VI. DISCUSSION

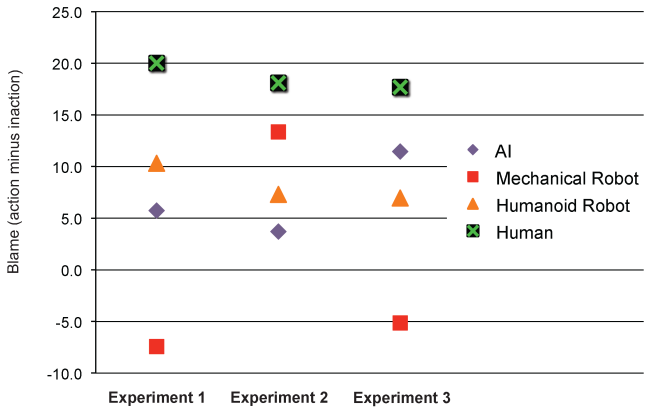
Previously we discovered a Human-Robot (HR) asymmetry in moral judgments [10]: that people blame robots more for inaction than action in a moral dilemma but blame humans more for action than inaction in the identical dilemma (where inaction allows four persons to die and action sacrifices one to save the four). In the present studies, we set out to examine whether this HR asymmetry still holds when the robot agent is depicted with a specific appearance and whether the HR asymmetry varies as a function of this appearance. In two out of three studies, we were able to demonstrate that the HR asymmetry indeed holds, but only when people make judgments about a mechanical-looking robot, not about a humanoid-looking robot. Patterns of blame for humanoid robots were very similar to those for human agents.

In general, our exploration of illustrations accompanying narrative experimental stimuli showed few variations (e.g., in picture format or number). But it did show that *identical* descriptions of a robot facing a moral dilemma can lead to different moral judgments of the robot's decision if the robot is portrayed as either a mechanical or a humanoid robot. The display of a mechanical agent may have triggered a mental model of robots as more rational, more "utilitarian," less affected by guilt or social reputation; and therefore people considered it morally less blameworthy when it sacrificed one life for the good of many. It would require detailed cultural studies to explore where such representations come from and how a single illustration can trigger such a rich representation. But the data suggest that, indeed, people treat a mechanical-looking robot differently from an identically-described human looking robot.

The HRI literature has shown that robots with human-like appearance are generally viewed as more agent-like, intelligent, and autonomous [37]. It seems more surprising to learn that moral judgments of a robot's actions can be altered by minor appearances. But before we draw too strong conclusions, we must consider the present studies' limitations.

One limitation is Experiment 2's failure to replicate the findings of Experiment 1. Researchers are increasingly becoming aware of the impossibility to consistently replicate past findings, especially with the same strength of effect [35]. Furthermore, the effect size of the moral HR asymmetry is relatively small and therefore more vulnerable to occasional replication failures. But plotting the findings of all three studies in Figure 6 does elicit some confidence in the general pattern of results.

Fig. 6. Difference scores of blame judgments for action vs. inaction across three experiments and four agent types.




At the same time, Figure 6 also highlights that people's blame judgments of the "AI" were similar to those of the humanoid robot and the human agent. Without further empirical evidence, we don't have a ready explanation for these results. We suspect, however, that the basis for blaming the AI more for action than inaction is different from the basis for blaming the humanoid robot in this way. Many people have some mental model of a humanoid robot, nourished by science-ficition literature and movies. This model may actually trigger the social-cognitive concepts and mechanisms that are normally conducive to ascribing moral blame to an agent. By contrast, most people do not have a mental model of an "AI" and may not know how to evaluate its moral decisions. In the absence of such a model, participants may resort to a generic baseline of what would be right or wrong in general (which is of course strongly influenced by their moral judgments of human agents).

A second limitation of our studies is the use of only one participant population, namely Amazon Mechanical Turk (AMT) contributors. These participants are more representative of the general population than typical student samples, but their greater heterogeneity in education, experience, and interest also leads to greater variability of responses. In the present studies we have assessed possible individual difference variables, such as demographics as well as general and specific perceptions of robots, but we found no impact on blame judgments; in other studies we also examined religious and political attitudes and interest in science fiction and robotics but found no moderating effects. The only variable that reliably moderates our findings is the willingness to treat robots as targets of moral judgments. Those who do grant robots such moral status show the HR asymmetry quite reliably, whereas those who don't tend to give blame ratings of 0 for what they consider to be mere machines programmed by humans and therefore appropriate targets of blame.

We also need to exercise caution in drawing too strong conclusions from our data because we have found surprising sensitivity of AMT samples to subtle variations in stimulus characteristics, such as reference to the potential victims of a moral dilemma situation and type of moral judgment (wrongness vs. blame). Moreover, in recent studies we have found effects of the order of judgments, not just for robot targets but for human targets. All of these variations remind us that we are at a very early stage of "moral HRI" [10] and that we must be prepared to find that some HRI results will vary by demographics, personality, experience, and cultural-historical changes.

Despite these caveats, our results do raise an important, and familiar question for robot design: Do we really want robots to look like humans and be treated like humans if they do not nearly have human-like capacities? For it appears that even moral judgments may be influenced by a robot's human-like appearance. Robot designers get to control what signals the robot emits to people who interact with it. However, if the signal does not match the capability, then sooner or later predictions on the human side will fail, expectations will be disappointed, and interactions with the robot will deteriorate. It may be particularly problematic to accept the risk of deceiving people about the robot's moral faculties, for false predictions of such facilities might end up causing significant personal and social harm. That is especially true in situations of life and death—the very ones that our experiments have begun to model.

## ACKNOWLEDGMENT

This project was supported in part by a grant from the Office of Naval Research, No. N00014-13-1-0269. The opinions expressed here are our own and do not necessarily reflect the views of ONR. We thank Justin Finkenaur for his patient and artful design of our pictorial stimuli.

## REFERENCES

[1] D. Victor, "Elon Musk and Stephen Hawking among hundreds to urge ban on military robots," 27-Jul-2015. [Online]. Available: http://www.nytimes.com/2015/07/28/technology/elon-musk-and-stephen-hawking-among-hundreds-to-urge-ban-on-military-robots.html?_r=0. [Accessed: 06-Oct-2015].

[2] B. F. Malle and M. Scheutz, "Moral competence in social robots," in *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics'2014*, Chicago, IL: IEEE, 2014, pp. 30–35.

[3] M. Scheutz and B. F. Malle, "'Think and do the right thing': A plea for morally competent autonomous robots.," in *Proceedings of the IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics'2014*, Red Hook, NY: Curran Associates/IEEE Computer Society, 2014, pp. 36–39.

[4] R. Sparrow, "Killer robots," *J. Appl. Philos.*, vol. 24, pp. 62–77, 2007.

[5] P. M. Asaro, "A body to kick, but still no soul to damn: Legal perspectives on robotics," in *Robot ethics: The ethical and social implications of robotics*, P. Lin, K. Abney, and G. Bekey, Eds. MIT Press, 2012, pp. 169–186.

[6] R. C. Arkin, *Governing lethal behavior in autonomous robots*. Boca Raton, FL: CRC Press, 2009.

[7] A. Litoiu, D. Ullman, J. Kim, and B. Scassellati, "Evidence that robots trigger a cheating detector in humans," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15*, 2015, pp. 165–172.

[8] W. Ju, *The design of implicit interactions*. Morgan & Claypool Publishers, 2014.

[9] R. Buchanan, "Wicked problems in design thinking," *Des. Issues*, vol. 8, no. 2, pp. 5–21, Apr. 1992.

[10] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many? People apply different moral norms to human and robot agents," in *HRI '15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.*, New York, NY: ACM, 2015, pp. 117–124.

[11] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen, "The neural bases of cognitive conflict and control in moral judgment," *Neuron*, vol. 44, no. 2, pp. 389–400, Oct. 2004.

[12] G. Briggs and M. Scheutz, "How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress," *Int. J. Soc. Robot.*, vol. 6, no. 2, pp. 1–13, 2014.

[13] 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.

[14] 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," *Conscious. Cogn.*, vol. 27, pp. 100–108, Jul. 2014.

[15] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, "All robots are not created equal: The design and perception of humanoid robot heads," in *Proceedings of the 4th Conference on Designing Interactive Systems (DIS '02): Processes, Practices, Methods, and Techniques*, 2002, pp. 321–326.

[16] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel, "'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," *Proc. 7th ACMIEEE Int. Conf. Hum.-Robot Interact. HRI12*, pp. 125–126, Mar. 2012.

[17] J. Forlizzi, "Towards the design and development of future robotic products and systems," p. 506, Aug. 2007.

[18] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," *Proc. Tenth Annu. ACMIEEE Int. Conf. Hum.-Robot Interact. HRI 15*, pp. 51–58, 2015.

[19] P. J. Hinds, T. L. Roberts, and H. Jones, "Whose job is it anyway? A study of human-robot interaction in a collaborative task," *Hum.-Comput. Interact.*, vol. 19, no. 1–2, pp. 151–181, Mar. 2004.

[20] C. Harrison, "Visual social semiotics: Understanding how still images make meaning," *Tech. Commun.*, vol. 50, no. 1, pp. 46–60, Feb. 2003.

[21] J. D. Bransford and M. K. Johnson, "Contextual prerequisites for understanding: Some investigations of comprehension and recall," *J. Verbal Learn. Verbal Behav.*, vol. 11, no. 6, pp. 717–726, Dec. 1972.

[22] P. N. Johnson-Laird, *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press, 1983.

[23] A. Mani and H. Sundaram, "Modeling user context with applications to media retrieval," *Multimed. Syst.*, vol. 12, pp. 339–353, Aug. 2006.

[24] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A theory of blame," *Psychol. Inq.*, vol. 25, no. 2, pp. 147–186, 2014.

[25] J. Voiklis, C. Cusimano, and B. F. Malle, "A social-conceptual map of moral criticism," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, P. Bello, M. Guarini, M. McShane, and B. Scassellati, Eds. Austin, TX: Cognitive Science Society, 2014, pp. 1700–1705.

[26] 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*, vol. 293, no. 5537, pp. 2105–2108, Sep. 2001.

[27] E. M. Caruso and F. Gino, "Blind ethics: Closing one's eyes polarizes moral judgments and discourages dishonest behavior," *Cognition*, vol. 118, no. 2, pp. 280–285, Feb. 2011.

[28] E. Amit and J. D. Greene, "You see, the ends don't justify the means: Visual imagery and moral judgment," *Psychol. Sci.*, vol. 23, no. 8, pp. 861–868, Aug. 2012.

[29] S. McCloud, *Understanding comics: The invisible art*. New York, NY: HarperPerennial, 1994.

[30] 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*, vol. 8, no. 3, p. e57410, Mar. 2013.

[31] W. Mason and S. Suri, "Conducting behavioral research on Amazon's Mechanical Turk," *Behav. Res. Methods*, vol. 44, pp. 1–23, Mar. 2012.

[32] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgm. Decis. Mak.*, vol. 5, no. 5, pp. 411–419, 2010.

[33] G. Paolacci and J. Chandler, "Inside the turk understanding mechanical turk as a participant pool," *Curr. Dir. Psychol. Sci.*, vol. 23, no. 3, pp. 184–188, Jun. 2014.

[34] F. L. Schmidt, "Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers," *Psychol. Methods*, vol. 1, no. 2, pp. 115–129, 1996.

[35] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *Science*, vol. 315, no. 5812, pp. 619–619, Feb. 2007.

[36] F. Eyssel and D. Kuchenbrandt, "Social categorization of social robots: Anthropomorphism as a function of robot group membership," *Br. J. Soc. Psychol.*, vol. 51, no. 4, pp. 724–731, Dec. 2012.

[37] A. Powers, A. D. I. Kramer, S. Lim, J. Kuo, S. Lee, and S. Kiesler, "Eliciting information from people with a gendered humanoid robot," *IEEE Int. Workshop Robot Hum. Interact. Commun. ROMAN 2005*, pp. 158–163, Aug. 2005.