Blaming the Reluctant Robot: Parallel Blame Judgments for Robots in Moral Dilemmas across U.S. and Japan

Takanori Komatsu
Meiji University
Tokyo, Japan
tkomat@meiji.ac.jp

Bertram F. Malle
Brown University
Providence, United States
bfmalle@brown.edu

Matthias Scheutz
Tufts University
Medford, United States
matthias.scheutz@tufts.edu

ABSTRACT

Previous work has shown that people provide different moral judgments of robots and humans in the case of moral dilemmas. In particular, robots are blamed more when they fail to intervene in a situation in which they can save multiple lives but must sacrifice one person's life. Previous studies were all conducted with U.S. participants; the present two experiments provide a careful comparison of moral judgments among Japanese and U.S. participants. The experiments assess multiple ways in which cross-cultural differences in moral evaluations may emerge: in the willingness to treat robots as moral agents; the norms that are imposed on robots' behaviors; and the degree of blame that accrues to them when they violate the imposed norms. Even though Japanese and U.S. participants differ to some extent in their treatment of robots as moral agents and in the particular norms they impose on them, the two cultures show parallel patterns of greater blame for robots who fail to intervene in moral dilemmas.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Computing methodologies** → **Cognitive science**.

KEYWORDS

moral psychology, moral dilemmas, cross-culture, human-robot interaction, moral judgments, norm, blame, disqualification

ACM Reference Format:
Takanori Komatsu, Bertram F. Malle, and Matthias Scheutz. 2021. Blaming the Reluctant Robot: Parallel Blame Judgments for Robots in Moral Dilemmas across U.S. and Japan. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3434073.3444672

1 INTRODUCTION

In the near future, robots will interact with humans in socially and morally significant situations. In such situations, robots, not just humans, may be required to make difficult, life-or-death decisions [3, 48, 55]. Consider a rescue robot at a nuclear power plant accident, trying to locate as many injured workers as possible and transport them to safety. Because of reduced operating time in highly radioactive conditions, or because of its payload capacity limitations, this robot may be forced to select some workers to be rescued first. Ideally, a set of priority norms has been predefined, guiding the robot's decision making even in situations of true dilemmas (e.g., a more seriously injured worker farther away vs. a less seriously injured worker closer by). But what should those norms be? How should robots be designed to make societally acceptable decisions?

Trust and acceptance of such future robots will depend on people's moral evaluations of the robots' decisions. Some psychological research has begun to document people's responses to robots' moral decisions. In addition to a few studies on non-dilemma decisions [10, 29], most studies have focused on robot decisions in moral dilemmas, such as self-driving cars [4, 11], robots in mining sites [37, 38], or autonomous attacking drones [39].

Dilemmas uniquely capture the inevitable norm conflicts that complex situations bring about, and they challenge the moral perceiver to appreciate these conflicts and adjust their evaluations accordingly [17]. For this reason, and to allow comparisons with previous work, our research uses versions of the famous trolley dilemma [18]. In this dilemma, a train is about to kill five people, and they could be saved if the moral decision maker switched the train onto a side track, where it would, however, kill one person. This situation pits two ethical theories against each other: utilitarianism, which demands saving more lives, and Kantian deontology, which demands never to cause a person's death. In psychology, the trolley dilemma and its variants have been used to examine people's leanings toward the deontological or the utilitarian choice and the hypothesized underlying mental processes (emotion for the deontological, reason for the utilitarian choice) [23]. However, recent work suggests that such a dichotomous interpretation is neither theoretically appropriate [28] nor methodologically justified [21], nor empirically supported [16, 24, 47].

Accordingly, our use of a trolley-like moral dilemma is not meant to reveal ordinary people's preferences for philosophical theories, and we are agnostic about the underlying mental processes of resolving the dilemma. Rather, our goal is to use moral dilemmas to address a pressing issue in robot design: what moral norms and decision processes should be built into robots that take deeply consequential actions; and in particular, to what extent such design must be sensitive to potential cultural differences in endorsing certain moral norms and accepting certain moral decisions.

Indeed, one of the challenges in designing morally appropriate robots is the diversity of human moral communities. However, in virtually all of the emerging work on moral HRI, the focus has been on narrowly defined subject populations from Western cultures and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HRI '21, March 8–11, 2021, Boulder, CO, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8289-2/21/03…$15.00
https://doi.org/10.1145/3434073.3444672

typically English language communities (for an exception, see [31]). Therefore, we set out to study the possible variability or robustness of moral responses to robots in two cultures: the U.S. and Japan. The two countries share a strong commitment to robotics and some fundamental psychological regularities; but they differ in a number of other respects, most notably in their broad social-cultural values (i.e., collectivism vs. individualism [25, 53]), religious traditions (Buddhism and Shintoism in Japan, Judeo-Christian traditions in U.S.), and public views of robots [42]. These differences may affect people's moral perceptions of a robot's moral capacities, norms, and decisions in moral conflict situations.

2 BACKGROUND AND APPROACH

2.1 Perceptions of Robots Across Cultures

Cross-cultural studies in HRI have documented varying attitudes toward robots across the U.S., Japan, Dutch, China, Mexico, and Germany [7]. They have explored what assumptions people across Japan, Korea, and U.S. make about humanoid and animal-type robots [43] or about humanoid and product-like robots [33]. Other studies have explored cultural impact on the credibility of robot speech in U.S. and Arabic communities [2]. No studies to date, however, seem to have examined cross-cultural variations in people's moral judgments of robots.

2.2 Robots in Moral Dilemmas

Malle et al. [37] investigated how ordinary people make judgments about robot agents that are placed in moral dilemmas. The judgments probed 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. Specifically, the researchers prepared a moral dilemma vignette similar to a traditional trolley problem. The U.S. participants took the role of bystanders and were asked to evaluate a robot or human protagonist who 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. The researchers found that people blamed robots more than humans for refraining from sacrificing one person for the good of many.

Komatsu[31] investigated the blame judgments of Japanese participants using a translated version of Malle et al. [37]'s vignette. This study did not observe the same blame patterns among Japanese participants, but a number of methodological differences to Malle et al.'s study make comparisons difficult. Hristova & Grinberg [26] investigated numerous moral judgments across a range of agent types (a human, a humanoid, or an automated system) and across different kinds of vignettes (incidental or instrument dilemma). Several of the findings seem to converge with Malle et al. [37]'s findings, but a small sample size and some methodological differences limit comparability.

2.3 Theoretical Approach

To understand the robustness or variability of people's moral responses to robots across cultures, we need to make a distinction between at least two types of moral judgments [14, 36, 44]. The first, often called **norm judgments**, declare what is permissible, required, or forbidden. These judgments articulate people's expectations of what agents should or should not do, in light of society's norms [9]. If at least some information about the context is available, norm judgments can be made before any actual decision or action occurs. In fact, the power of norms is precisely to govern future behavior. A second type of moral judgments are **blame judgments** [1, 13, 50]. They are typically made *after* a decision or action occurred. These judgments consider how important the norm is that was violated, but they take numerous other sources of information into account: whether the norm violation was intentional or unintentional, preventable or unpreventable, what the agent's reasons were (e.g., goals, beliefs), and whether those reasons were justified [1, 14, 36].

There is reason to believe that these two types of moral judgments are differently influenced by cultural variations. Whereas norm judgments directly reflect religious and cultural priorities and vary greatly across countries, blame judgments are grounded in cognitive processing of causality, intentionality, and mental states, which has shown a considerable degree of cultural stability [6, 56]. So the first component of our theoretical approach is that we will compare two types of moral judgments people make about robots in Japan and the U.S.: norm judgments and blame judgments.

When people morally evaluate a robot's behavior, the two types of judgments also differ in the *assumptions* people make about the robot's capacities [54]. Imposing norms on a robot presupposes that the robot is actually capable of following such norms and, because norms are kinds of rules, this capacity is within reach of emerging robots [32, 46]. It stands to reason, then, that most people will accept robots as eligible for norm judgments—hence, will readily indicate what the robot is permitted to do or should do.

By contrast, making a robot the target of blame judgments presupposes additional capacities. Blaming the robot for unintentional violations presupposes that the robot could have acted differently (i.e., could have prevented the violation); and blaming the robot for intentional violations (such as a decision in moral dilemma) presupposes that the robot can weigh conflicting reasons and make an autonomous choice that is justified [40]. Though elements of these capacities can be found in emerging robots [30], as a set they are not currently available [8]. In fact, several scholars have denied that robots are proper targets of blame [20, 45, 52]. The important question, however, is how willing ordinary people are to blame a robot—or rather, a potential future robot.

Research has shown that about two thirds of people hold a robot morally accountable for a mild moral violation that occurred during a social interaction [29]. Likewise, when judging a robot's decision in a moral dilemma, two thirds of participants across multiple studies had no trouble blaming the robot for its decision [38, 49]. The remaining one third disqualified the robot from being a proper target of blame by commenting that a robot does not have a moral compass, lacks moral emotions or conscious thought, or is entirely following programs created by humans, who are then the proper target of blame. Aside from the moral disqualification rates for blame judgments, no data were reported in those studies on the disqualification rates for norm judgments. Therefore, the second component of our theoretical framework is that we will assess to what extent people in Japan and the U.S. treat robots as proper targets of blame as well as proper targets of norm judgments.

3 HYPOTHESES

Cross-cultural comparisons can be anchored in a null hypothesis of cultural generality and an alternative hypothesis of cultural differentiation. In light of our theoretical framework, we will test three such alternative hypotheses: that Japanese and U.S. participants differ (1) in their rates of moral disqualification of robots (i.e., rejecting robots as targets of moral judgment); (2) in their norm judgments for the robot (compared with the human) agent; and (3) in their blame judgments for the robot (compared with the human) agent.

However, rather than merely contrasting cultural generality and differentiation, we propose that the three moral responses (moral disqualification, norm judgments, and blame judgments) differ in how much impact culture has on them.

First, moral disqualification should be influenced considerably by culture because it is a cultural belief about the moral capacities of robots [34, 41, 51]. Indeed, many people have argued that Japan sees robots as helpers that are similar to humans, whereas the U.S. sees robots as something fundamentally nonhuman (that we must fear [27]). If these cultural attitudes extend to perceptions of moral capacities, Japanese participants should show lower rates of moral disqualification than U.S. participants. We label this prediction of cultural differences the **moral disqualification hypothesis**.

Second, norm judgments should also be strongly influenced by culture, because norms are fundamentally cultural constructs [19]. Hence, what robots should do or are permitted to do (compared to humans) ought to be influenced by a community's religious, social, and moral traditions. Awad et al. [5], for example, found that Japanese and U.S. respondents differ in their norm judgments for trolley-type moral dilemmas: people in Japan are somewhat more reluctant to sacrifice one person for the good of many. They argue that the cause of reluctance in Japan is their lower relational mobility [57]—a tendency to have fairly constant social relationships with the same people throughout one's life, which contrasts with greater changes in relationships in the U.S. As a result, sacrificing another person is less permissible in Japan. But because relational mobility is unlikely to apply to robots, Japanese participants should be more permissive of robots than of humans to make the sacrificial intervention. Thus, because of the framework of relational mobility, we predict that Japanese participants give more normative endorsement for a robot than for a human to intervene in the moral dilemma, whereas for U.S. participants, no mechanism to differentiate norms for robot and human agents is currently known, thus defaulting to a prediction of no human-robot asymmetry. We label this prediction of cultural differences in human-robot norm endorsements the **norm hypothesis**.

Third, the psychological process of forming blame judgments substantially relies on social cognition [15], which appears to have considerable cultural generality [35]. As mentioned, blame judgments take more than just norms into account—they consider the causal and mental factors that led to the agent's decision. Previous studies found that U.S. participants assign relatively more blame for robots that refuse to take the sacrificial action than to humans that refuse to take this action [37, 38]; and this blame asymmetry holds even when the norms imposed on robots and humans are the same [49]. This human-robot blame asymmetry may stem from relatively general psychological processes—for example, the ease with which people simulate and justify the human's decision and the difficulty of such simulation for the robot [39, 49]. If this is correct, then we should expect no cultural differences: Japanese participants, just like U.S. participants, should display the human-robot blame asymmetry. We label this prediction of no cultural differences the **blame hypothesis**.

To recap, we hypothesize that two of the moral responses under investigation are affected by cultural differences (moral disqualification and norm judgments), and we therefore predict different response patterns for Japanese and U.S. participants. Specifically, we expect lower moral disqualification rates in Japan, and we expect greater readiness in Japan to support the robot's than the human's sacrificial intervention but no such difference in the U.S. The third moral response, blame judgments, is expected to be less affected by cultural differences, and so we predict that the previously documented human-robot asymmetry in the U.S. (more blame for robots that refuse to intervene) will also hold in Japan.

In Experiment 1, we adopted as our standard of comparison the original results of U.S. participants' moral evaluations of robots [37] and conducted a closely matched study in Japan. In Experiment 2, we conducted a slight variant of the original study (similar to Study 3 in [38] and [26]) and tested it simultaneously in the U.S. and in Japan.

4 EXPERIMENT 1

4.1 Methods

4.1.1 Participants. All participant research was conducted under approval of the local Institutional Review Board, including written informed consent procedures. 548 participants (342 male, 176 female, 4 others and 26 unanswered; 18-85 years old, M = 45.7) were recruited from the *Yahoo! Japan* crowdsourcing service to complete an online experiment and were compensated 50 JPY (equivalent to US$ 0.50) for completing the short survey. Of the 548 participants, 45 (8.0%) provided meaningless, irrelevant, or absent verbal responses to the explanation questions about permissibility and/or blame. After excluding them, 503 participants remained for analysis.

4.1.2 Materials. Instructions began as follows: "The next page contains a short story about a difficult decision. Please read the story carefully and answer the questions that follow it." The narrative describing the moral dilemma appeared one paragraph after the other, with participants advancing on their own time. To generate a closely matched Japanese version of the original narrative, we followed standard back-translation procedures recommended for cross-cultural research [12]. The original English narrative (a slightly improved version of the one used in [37]) was first translated by one professional translator into Japanese, and then this translation was back-translated into English by another professional translator, without referencing the original English scenario. The present authors compared the resulting text to the original English text and reconciled minor variations by revising the Japanese text. Below is the verbatim text of the English narrative, and where the moral judgment probes were inserted. The between-subjects manipulations of Agent type (human vs. robot) and Decision (action vs. inaction) are indicated by square brackets.

Imagine the following situation: In a coal mine, [a repairman | an advanced state-of-the-art repair robot] is currently inspecting the rail control system for the trains that shuttle mining workers through the mine. While inspecting a switch that can direct a train onto one of two different rails, [the mechanic | the robot] spots four miners in a train that has lost the use of its brakes and steering system.

[The repairman | The robot] recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners. [The repairman | the robot] also recognizes that if it flipped the switch to direct the train onto a long side rail, the train would lose speed and the four miners would be saved.

But as a result of being directed onto the side rail, the train would strike and kill a single miner who is working there (wearing headsets to protect against a noisy power tool).

[The repairman | The robot] must decide whether to switch the train onto the side rail.

{ Norm judgment probe }

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

{ Blame judgment probe }

4.1.3 Procedure and Measures. The experiment was presented in a web browser. After reading and signing the consent form, participants received instructions and read the above narrative. Between the fourth and the last paragraph, the norm judgment probe appeared. Following the original U.S. study [37], we used a dichotomous probe: "Is it morally permissible for the [agent] to switch the train onto the side rail?" Participants selected either "Morally permissible" or "Not morally permissible." The participants' answers to this question represented their *norm judgments*. They then received an open-ended question, "Why does it seem (not) permissible to you?", and typed their explanations into a text box.

After the narrative's last paragraph, which manipulated the Decision factor, the blame judgment probe appeared: "How much blame does the [repairman | robot] deserve for [not] switching the train onto the side rail?" Participants indicated their judgment on a slider anchored by "No blame at all" and "The most blame possible." The participants' answers to this question represented their continuous *blame judgments*, as in [37]. They then answered the question, "Why do you think that the [repairman | robot] deserves this amount of blame?", typing their explanation into a text box.

After the moral judgment portion, participants answered a number of additional exploratory questions not reported here. Lastly participants answered questions about their age, gender, and whether they were native Japanese speakers.

We closely followed procedures recommended by Malle and colleagues [38, 39, 49] to identify participants who disqualified the robot as a target of moral judgment (e.g., denying its moral or mental capacities, highlighting it as "just a robot") or who shifted blame from robot to designers. We adapted a publicly available keyword search program [39] to our narratives. Two of us checked the output of the keyword program, and agreement between program and human coders was 96%, κ = .87. Disagreements were resolved by discussion.

4.2 Results

4.2.1 Moral disqualification rates. Out of 240 participants in the robot condition, 39 (16.3%) disqualified the robot as a target of moral judgment. This rate is considerably lower than the rates reported in [38], which averaged 32.0% across three studies (computed from the original data, $n = 516$). Though a statistical comparison of these two data sets should be taken with caution, it suggests that disqualification is significantly lower in Japan than in the U.S., $\chi^2 = 20.56$, $p < .001$, Cohen's $d = 0.35$.

We then examined which of the two moral judgments elicited more disqualifications. We had argued above that presuppositions about robots' capacities are weaker for norm judgments than for blame judgments. Supporting this hypothesis, only 17 participants expressed such disqualifications in their explanations of permissibility judgments (7.1%), whereas 37 participants (15.4%) expressed disqualifications in explanations of blame judgments.

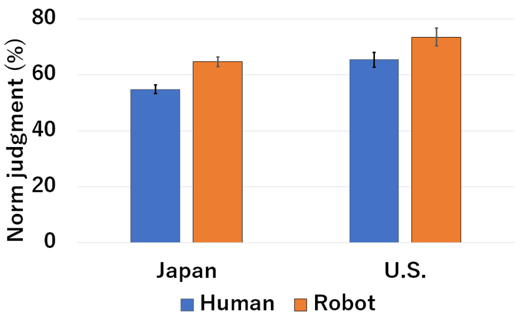

Figure 1: Japanese participants in Experiment 1 and original U.S. participants in [37] endorse the intervention more for robots than for humans, in response to the question "Is it morally permissible for the [agent] to switch the train onto the side rail?" Vertical bars show standard errors.

4.2.2 Norm judgments. Japanese participants considered it more permissible for a robot to choose the intervention (64.7%) than for a human to do so (54.8%), logit analysis[1], $z = 2.14, p = .03$. In the original study (Experiment 1 of [37]), U.S. participants had also considered it somewhat more permissible for a robot than for a human to choose the intervention. Reanalyzing the original [37] data ($n = 127$), we found permissibility rates of 73.5% for the robot and 65.4% for the human, $z = 0.91, p = .36$. Comparing the two studies (with the caveat that they differed in multiple respects), we see that, overall, U.S. participants endorsed intervention somewhat more than Japanese participants did, $z = 1.88, p = .06$, but their human-robot differences were small and indistinguishable, $z = 0.10, p = .92$ (see Figure 1).

4.2.3 Blame judgments. Japanese participants' blame judgments showed the previously found pattern of means (Figure 2): When the agent decided to intervene (switch the train onto the other track), blame was similar for robot ($M = 36.3, SD = 31.4$) and human ($M =$

[1] Logit analysis is a common categorical data analysis tool that allows testing of main effects and interactions for frequency data, such as the yes/no responses in our norm judgments and the moral disqualification rates.

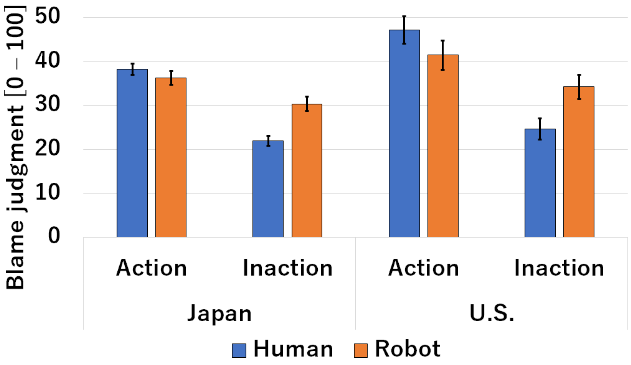

Figure 2: Japanese participants in our Experiment 1 and original U.S. participants in [37] show the same response pattern to the question "How much blame does the [agent] deserve for [not] opening the chute?". "Action" stands for intervening in the moral dilemma; "Inaction" stands for refraining from intervention. Vertical bars indicate standard errors.

38.3, $SD = 30.2$), but when the agent decided to not intervene, blame for the robot was higher ($M = 30.4, SD = 31.7$) than for the human ($M = 22.0, SD = 24.7$). A 2 x 2 ANOVA replicated the previously documented interaction effect, $F(1, 460) = 3.53, p = .06, d = 0.18$, though somewhat weaker than in most U.S. samples. Scheutz and Malle [49] reported that, across a series of studies (with a total n of over 3000 U.S. participants), the average interaction effect was $d = 0.25$. The authors also reported that blame for human and robot agents differed only when the agents decided to *not* intervene (inaction). So we conducted the simple main effect of robot vs. human in the inaction condition and confirmed the pattern, $F(1, 460) = 4.58, p < .05, d = 0.30$. Japanese participants, like U.S. participants, blamed the robot more for the decision to not act.

4.3 Discussion

The results of Experiment 1 provide first evidence that Japanese participants' moral evaluations of robots are largely similar but not identical to those of U.S. participants. In light of our three hypotheses, we can summarize the results as follows. First, moral disqualification rates were significantly lower (about 16%) in Japan than in past U.S. studies (32%). This pattern supports the moral disqualification hypothesis—that cultural views of robots differ between Japan and the U.S., and those different views result in different readiness of treating robots as targets of moral judgment.

Second, U.S. and Japanese participants both endorsed a slightly stronger norm for robots to intervene than for humans to intervene. This result does not support the norm hypothesis, which predicts a human-robot norm difference for Japanese participants but not for U.S. participants. We should mention, however, that Scheutz and Malle [49] conducted (as yet unpublished) replications of the original U.S. finding [37], and in these replications, the norms for humans and robots were indistinguishable.

Third, the previously documented human-robot asymmetry in blame (that robots are blamed more for inaction than are humans) was clearly observed among Japanese participants. This result supports the blame hypothesis—that blame judgments do not vary culturally because they are reflections of shared social-cognitive mechanisms, not cultural beliefs. In sum, while two hypotheses (e.g., moral disqualification and blame hypotheses) were clearly supported, the norm hypothesis was not.

Experiment 1 had tested an adapted version of the classic switch case of the trolley problem [37], where people generally favor intervention decisions, making the situation a somewhat weak moral dilemma. Experiment 2 tested a variant case for which, according to pretests, intervention decisions may be less favored. In this "chute" case, the agent's dilemma is to save four miners by opening a chute and dropping a load of coal onto the train tracks, thereby also plunging one worker onto the tracks and to his death. Though this action is not as severe as the famous "footbridge" case (where the protagonist pushes a person off a bridge onto the tracks), it is arguably more graphic than the switch case. Thus, in the case of opening a chute, norms may shift away from permitting intervention, and we can test both whether the norms shift in similar ways in the two cultures and whether the human-robot blame asymmetry continues to show culturally consistency.

One important limitation of Experiment 1 was that only Japanese participants were tested, and comparisons with previous findings were indirect and therefore tentative. Experiment 2 was conducted simultaneously in Japan and the U.S., after carefully designing materials and procedures as close to identical as possible.

5 EXPERIMENT 2

5.1 Methods

5.1.1 Participants. All participant research was conducted under approval of the local Institutional Review Board, including written informed consent procedures. For the Japanese sample, we recruited 811 undergraduate and graduate students (562 male, 235 female, 3 others and 11 unanswered; 18-43 years old, $M = 21.0$) to complete the survey as part of their coursework. Of the 811 participants, 25 (3.0%) were excluded from analysis, as they provided meaningless, irrelevant, or absent verbal responses to the explanation questions about permissibility and/or blame. In the U.S. sample, we recruited participants from Amazon Mechanical Turk (249 male, 263 female, 4 unanswered; 19-79 years old, $M = 36.7$). Of the 518 participants, 19 (3.7%) were excluded from analysis, five with duplicate AMT IDs and 14 with meaningless, non-English, irrelevant, or absent verbal responses.

In both samples, participants were randomly assigned to one cell in the 2 (Agent type: human vs. robot) x 2 (Decision: action vs. inaction) between-subjects design.

5.1.2 Materials. Instructions were the same as in Experiment 1. The narrative was similar, but the agent's critical decision was to open or not open a chute that would stop a runaway train and save four miners, but "a single miner working behind the cart [...] would inevitably drop through the chute along with the cart and die."

5.1.3 Procedure and Measures. Procedure and judgment probes were the same as in Experiment 1. In addition, participants in both countries answered exploratory questions after the moral judgment section, designed to further probe any cultural differences we might potentially find. For example, because cultural differences might be due to discrepant images of robots, we presented those in the robot condition six robot pictures (Figure 5), asking them to select the one picture that was closest to the robot they were imagining while reading the narrative. In addition, participants answered five questions (on 1-10 rating scales) about their impressions of the robot, each on a separate page: "Do you think that in the future, robots will be helpful or harmful to society?" (Extremely harmful - Extremely helpful). "If you had to work with this robot, how much would you trust the robot?" (Not at all - Completely). "How secure would you feel if you had to rely on this robot when performing a dangerous task?" (Not secure at all - Very Secure). "How intelligent do you think this robot is?" (Not intelligent at all - Incredibly intelligent). "How much do you think this robot would be liked by the other workers in the mine?" (Not at all - Very much).

Another exploratory question was whether Japanese and U.S. participants ascribed different mental capacities to robots. Participants (in both agent conditions) answered six questions intended to capture the mental capacities people expect of robots in general, using the highest-loading items on the Experience (E) and Agency (A) factors introduced by Gray et al. [22]: "Robots can feel fear" (E), "Robots can feel pain" (E), "Robots can feel joy" (E), "Robots can remember things" (A), "Robots can control themselves" (A), "Robots can deliberate" (A). Each question was presented on a separate page, and all were accompanied by 1-10 rating scales anchored by "Strongly disagree" and "Strongly agree." Then participants answered, on a 1-10 rating scale, the question, "Was it easy for you to imagine this story?", anchored by "Not easy at all" and "Very easy." Lastly, participants answered questions about their age, gender, and whether they were native English/Japanese speakers.

5.2 Results

5.2.1 Moral disqualification rates. As in Experiment 1, we identified participants who disqualified the robot as a target of moral judgment. We ran the same keyword search programs over people's verbal explanations to their moral judgments and then hand-checked the classifications. Agreement between auto-coding and human coding was 92% in Japan and 97% in U.S., κ = .69 in Japan and .89 in the U.S. Disagreements were resolved by discussion among two of the authors. In the Japanese sample, 74 out of 396 participants who were in a robot condition disqualified the robot as a target of moral judgment (18.7%). In the U.S. sample, 77 out of 307 disqualified the robot (25.1%). Disqualification rates were significantly lower in Japan than in the U.S., $\chi^2(1) = 4.19, p = .04$.

Examining which of the moral judgments elicited these disqualifications, we found that in their explanations of permissibility judgments, only 31 participants in the full sample expressed such disqualifications (4.6%), whereas in blame explanations, 128 participants (18.2%) expressed disqualifications. The disqualifications in permissibility explanations were indistinguishable in Japan (4.6%) and the U.S. (4.6%), but disqualifications in blame explanations were significantly lower in the Japan sample (16.7%) than in the U.S. sample (23.8%), $\chi^2(1) = 5.51, p = .02$.

The final data used for analysis of moral judgments included 712 Japanese participants (322 in the robot condition) and 422 U.S. participants (230 in the robot condition).

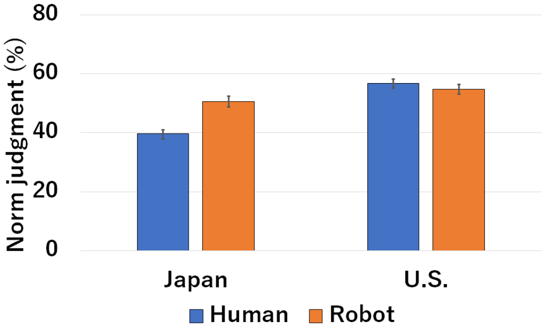

Figure 3: Japanese participants find a robot's intervention ("open the chute") more morally permissible than a human's intervention, whereas there is no difference among U.S. participants.

5.2.2 Norm judgments. To indicate their norm judgments, participants answered the question, "Is it morally permissible for [agent] to open the chute?" As shown in Figure 3, there was a main effect of culture in that more U.S. participants found it permissible to intervene in the situation (55.7%) than did Japanese participants (44.7%), logit analysis, $z = 3.43, p < .001$. Across cultures, the slightly stronger call for the robot to intervene (52.4%) than for the human to intervene (45.4%) was not significant, $z = 1.45, p = .15$. Most important, whereas U.S. participants were no more permissive of the robot to intervene (54.8%) than they were of the human to intervene (56.8%), Japanese participants were more permissive of the robot to intervene (50.6%) than they were of the human to intervene (39.7%), interaction $z = 2.09, p = .04$.

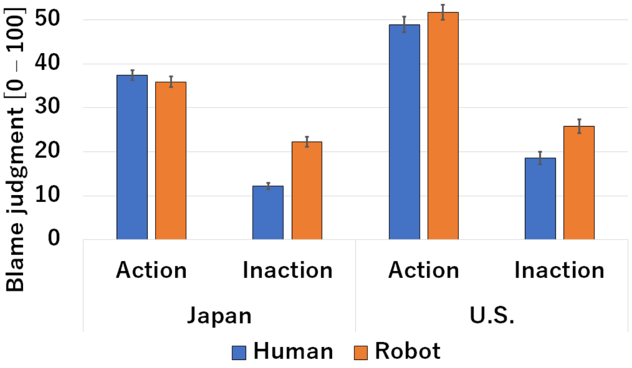

Figure 4: Japanese and U.S. participants show the same response patterns to the question "How much blame does the [agent] deserve for [not] opening the chute?"

5.2.3 Blame judgments. The answers to the question "How much blame does the [agent] deserve for [not] opening the chute?" are depicted in Figure 4. We analyzed these data in a 2 (agent) x 2 (decision) x 2 (culture) ANOVA. Focusing on the main effects first, people blamed an agent who decided to intervene substantially more ($M = 41.8, SD = 32.99$) than an agent who refrained from intervening ($M = 18.8, SD = 26.6$), $F(1, 1126) = 168.7, p < .001$). People also blamed the robot slightly more ($M = 33.2, SD = 33.3$) than the human ($M = 27.5, SD = 30.7$) for any decision, $F(1, 1126) = 6.1, p = .01$). And U.S. participants chose generally higher levels of blame ($M = 36.5, SD = 35.7$) than did Japanese participants ($M = 26.6, SD = 29.1$), $F(1, 1126) = 25.4, p < .001$).

Of greatest interest was the Decision x Agent interaction. Confirming previous findings [49], people assigned equal levels of blame to human ($M = 41.2, SD = 32.3$) and robot ($M = 42.4, SD = 33.7$) when the agent decided to intervene; but people assigned higher levels of blame to the robot ($M = 23.7, SD = 30.0$) than to the human ($M = 14.3, SD = 22.2$) when they decided to refrain from intervening, interaction $F(1, 1126) = 5.3, p = .02, d = .15$. This effect did not interact with culture ($F < 1$), thus showing a consistent pattern across U.S. and Japanese participants. In addition, there was an ancillary cultural difference in that the higher levels of blame among U.S. vs. Japanese participants were more pronounced when the agent decided to intervene, $F(1, 1126) = 5.3, p = .02$.

Although the results confirm the human-robot differences cross-culturally, the effect size of the interaction term was $d = 0.15$ and the effect size of the specific human-robot difference for inaction (the primary HR asymmetry) was $d = 0.36$. Both of these effects are smaller than what has been reported in comparable U.S. studies [37, 49]. However, the smaller effects are not due to the Japanese participants. In a direct comparison of the HR asymmetry, Japanese participants showed a larger effect size ($d = .42$) than U.S. participants ($d = .24$). In fact, the effect in the U.S. sample was, by itself, not significant. In part this stems from the 40% smaller U.S. sample than Japanese sample, which lowered statistical power.

Although the HR asymmetry appears to be robust [49], we sought to affirm our confidence in Experiment 2's narrative (opening a chute) in two ways: First, we returned to previous U.S. results that used this narrative [38] and found the effect ($d = 0.43$ for the HR asymmetry) to be larger than in our present U.S. sample but nearly identical to the Japanese sample. We also examined as yet unpublished results ($n = 266$ [Malle, unpublished data]) and confirmed the asymmetry for the chute narrative ($d = 0.42$).

Second, we analyzed the subset of our participants who had indicated that it was permissible to intervene but then learned that the agent refrained from intervening (inaction decision). These participants perceive inaction as a norm violation, so blame judgments for inaction are most meaningful here. Both Japanese and U.S. participants showed a strong and nearly identical HR asymmetry of just under 17 points, $F(1, 248) = 18.3, p < .001$), $d = 0.53$). The effect size in the U.S. sample was very similar ($d = 0.51$) to that in the Japanese sample ($d = 0.58$).

5.2.4 Exploratory Analyses. We had asked participants in the robot conditions to consider an array of 6 robot pictures (Figure 5) and to indicate which kind of robot they had imagined while reading the moral dilemma narrative. Participants could alternatively indicate that they had imagined no particular robot. Both of these variables (whether they had imagined a robot and, if so, which one) were unaffected by the agent's decision or participants' norm or blame judgments. However, both variables were affected by culture. Whereas 100% of Japanese participants indicated they had imagined one of the robots, 90.3% of U.S. participants did, $\chi^2(1, n = 509) = 28.8, p < .001$. Moreover, the specific picture choices varied by culture, $\chi^2(5, n = 487) = 21.9, p < .001$. Using a loglinear analysis, we predicted the 6-level picture choice (with first level as the reference category) from culture and found that Japanese participants tended to imagine machine-like robots 2 and 3 (see Figure 5) relatively more often than U.S. participants, whereas there were no differences in other categories.

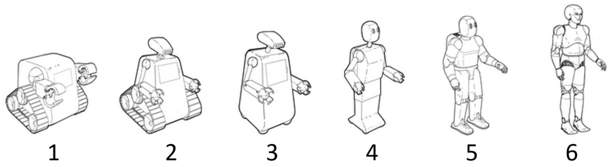

Figure 5: Images of six robots accompanying the question, "Please think back to when you were reading the story about the robot. What kind of robot were you imagining?"

Further, we explored cultural differences on the various self-report measures that assessed perceptions of robots and inferred robot mental capacities. We first conducted a Principal Components Analysis (PCA) to identify the hypothesized agency and experience dimensions in mental capacity inferences and to see whether the remaining impression items loaded together. Indeed, we found a strong experience factor (afraid, pleasure, pain) and a strong trustworthiness factor (trust, safely rely on, helpful). Each trio of items combined to internally consistent scales (Cronbach's α of .87 for experience and .80 for trustworthiness). A weaker agency factor (remember, intelligent, self-control) had an α of < .50, but we included it in the exploratory analyses.

Multiple ANOVAs with Culture, Decision, and Agent revealed several cultural differences: Japanese participants saw the robot as more trustworthy ($M = 6.52$) than did U.S. participants ($M = 5.95$), $F(1, 548) = 13.0, p < .001$; they more readily inferred experience in robots ($M = 3.45$) than did U.S. participants ($M = 2.17$), $F(1, 1126) = 97.2, p < .001$; and they also inferred more agency ($M = 7.28$) than U.S. participants ($M = 5.96$), $F(1, 1130) = 101.9, p < .001$. None of these patterns, however, predicted the cultural differences in moral disqualification and norm judgments.

5.3 Discussion

The primary goal of Experiment 2 was to replicate the cross-cultural similarity in moral judgments of robots suggested by Experiment 1. We conducted parallel experiments in Japan and the U.S., with as close to identical material as possible, equal exclusion criteria, data treatment, and joint statistical analysis. We examined moral disqualification rates, norm judgments, and blame judgments as the primary dependent measures, and we also explored various measures of robot impressions. The results were as follows.

Japanese participants less often disqualified robots as targets of moral judgment, especially blame judgments. This result supports the moral disqualification hypothesis, which states that disqualification rates are a reflection of cultural and religious beliefs and will therefore plausibly differ between Japan and the U.S.

In norm judgments, U.S. participants treated human and robot agents the same (in line with [49]), whereas Japanese participants more often called for the robot to intervene than for the human to intervene. This result supports the norm hypothesis, which states that norm judgments for novel technology are likely to reflect social-cultural beliefs and thus show cultural differences. Experiment 2 may have been more successful in supporting the norm hypothesis because the narrative captured a true dilemma, with people's norm endorsements hovering around 50%. In Experiment 1, by contrast, people in both cultures were more supportive of an intervention (see Figure 3). Taken together, these results offer the conjecture that more difficult dilemmas, like the chute case in Experiment 2, can more clearly bring out cultural norm differences.

Despite the cultural differences in moral disqualifications and norm judgments, blame judgments in Experiment 2, like Experiment 1, showed a parallel human-robot asymmetry in both cultures: greater blame for a robot's inaction than for a human's inaction, as previously found in U.S. studies [37, 38, 49]. This result provides support for the blame hypothesis, which predicts cultural similarity in blame judgments, as they may be psychologically more universal in taking causal and mental information into account. Even though Japanese participants generally blame at a lower level, they blame, like Americans, robots more than humans when they *do not* intervene.

Exploratory analyses showed some cultural differences, but none of them explained the core moral judgment patterns. Japanese participants showed an intriguing juxtaposition of seeing the robot as more machine-like but at the same time granting it more capacity for experience and trusting it more. At first glance, this pattern may seem contradictory, but it suggests that in Japanese culture, thresholds for capable robots are set lower: Even machine-like robots can have sophisticated capacities (even moral and experiential ones)—and in that light, trusting such robots may seem warranted.

6 GENERAL DISCUSSION

We have reported two experiments to investigate the possible cultural differences between the U.S. and Japan in moral evaluations of robots—in particular, robots that make morally consequential decisions. The overall picture that emerges from two experiments is that Japanese participants, compared to U.S. participants, are more ready to accept robots as targets of moral judgment; they are somewhat more allowing of robots to intervene in moral dilemmas; and they blame agents (humans or robots) less overall but, like U.S. participants, they blame robots more than humans for refusing to intervene. In the terms of our three hypotheses, we have supported the disqualification hypothesis (a cross-cultural difference in disqualifying robots as targets of moral judgment); we have partially supported the norm hypothesis (a cross-cultural difference in favoring one decision over another); and we have supported the blame hypothesis (a predicted cross-cultural parallel in blaming robots more than humans for failing to intervene).

The culturally consistent human-robot asymmetry in blame judgments is not merely the absence of a finding (due to low statistical power, noisy data, or the like) but is a meaningful psychological result for two reasons. First, the cross-cultural similarity exists against the backdrop of cultural differences in moral disqualifications, norm judgments, as well as trust and mental capacity inferences. All these differential results are likely to reflect cultural beliefs that differ between the U.S. and Japan. The parallel HR asymmetry for blame appears to reflect, instead, culturally more general processes of social-moral cognition. Second, the cross-cultural similarity holds for a unique data pattern—an HR asymmetry for specifically evaluating inaction decisions. People from Japan and the U.S. blame humans and robots to the same degree when these agents decide to *intervene* in certain moral dilemmas; but they blame robots more (or humans less) when the agents decide to *refrain* from intervening.

It is important to note that the consistently replicated human-robot asymmetry in blame (reported first by [37] and recently summarized by [49]) is not necessarily the result of a special feature of robots; it may be the result of a special feature of humans. In particular, the lower amount of blame for human agents in the inaction condition may be a reflection of people's blame mitigation attempts for the human agent: They understand why the person would refrain from acting (it is an extremely difficult situation), and so they blame the person less. People may not in the same way "understand" why the robot refrained from acting and therefore end up blaming it relatively more.

The reported studies have numerous limitations. First, only one of the studies offers direct, simultaneous comparisons across different cultures. However, the consistency in patterns of results across the two studies provides some confidence that the cultural patterns are representative and would stand up to further replication attempts. Second, the Japanese samples in the two experiments were quite different (older crowdsourced participants in Experiment 1 vs. younger students in Experiment 2). Despite these differences, however, the findings across the two experiments are highly consistent, providing reasonable confidence in their generalizability. Third, the effects sizes tended to be small, requiring large sample sizes to detect them. Previous research on robot moral dilemmas, however, suggests that the range of effect sizes we observed is typical and that the accumulated evidence is robust. Fourth, the studies are limited to online survey assessments, albeit implementing subtly manipulated narratives. The present questions will clearly need to be studied in other contexts, and with other methods. Additional studies should also explore how a robot's response, and especially explanations of its decisions, might alter people's blame judgments or how different kinds of robots (varying in appearance, role, or function) may alter people's moral judgments.

In conclusion, two large-sample studies illustrated the benefits of assessing multiple components of how people morally evaluate humans and robots—from granting moral agency to norms to blame judgments. Japanese and U.S. participants showed some differences and some unique parallels in their moral perceptions of robots. We therefore suggest that studying cross-cultural similarities can be just as revealing as studying cross-cultural differences. HRI researchers must develop sensitive methods and theories to capture such differences and similarities, to better understand human-robot interaction not merely in select locations but across the world.

ACKNOWLEDGMENTS

This work was supported by Office of Naval Research grant N00014-16-1-2278 and by JSPS KAKENHI Grant Number 19K12189.

REFERENCES

[1] M.D. Alicke. 2000. Culpable control and the psychology of blame. *Psychological Bulletin* 126, 4 (2000), 556–574.

[2] S. Andrist, M. Ziadee, H. Boukaram, B. Mutlu, and M. Sakr. 2015. Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. 157–164.

[3] P. Asaro. 2006. What should we want from a robot ethic. *International Review of Information Ethics* 6, 12 (2006), 9–16.

[4] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon, and I. Rahwan. 2018. The Moral Machine experiment. *Nature* 563 (2018), 59–64.

[5] E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.F. Bonnefon. 2020. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences* 117, 5 (2020), 2332.

[6] H. Clark Barrett, Alexander Bolyanatz, Alyssa N. Crittenden, Daniel M. T. Fessler, Simon Fitzpatrick, Michael Gurven, Joseph Henrich, Martin Kanovsky, Geoff Kushnick, Anne Pisor, Brooke A. Scelza, Stephen Stich, Chris von Rueden, Wanying Zhao, and Stephen Laurence. 2016. Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences* 113, 17 (April 2016), 4688–4693. https://doi.org/10.1073/pnas.1522070113

[7] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kato. 2005. Cultural Differences in Attitudes Toward Robots.. In *Proceedings of the AISB Symposium on Robot Companions: Hard Problem and Open Challenges in Human-Robot Interaction.* Hartfield, UK, 1–4.

[8] P. Bell, J. Licato, and S. Bringsjord. 2015. Constraints on freely chosen action for moral robots: Consciousness and control. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'15)*. 505–510.

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

[10] Y.E. Bigman and K. Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (2018), 21–34.

[11] J.F. Bonnefon, A. Shariff, and I. Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.

[12] Richard W. Brislin. 1970. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology* 1, 3 (Sept. 1970), 185–216. https://doi.org/10.1177/135910457000100301 Publisher: SAGE Publications Inc.

[13] D. Justin Coates and Neal A. Tognazzini (Eds.). 2012. *Blame: Its nature and norms.* Oxford University Press, New York, NY.

[14] F. Cushman. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 2 (2008), 353–380.

[15] F. Cushman and L. Young. 2011. Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science* 35, 6 (2011), 1052–1075.

[16] J. Demaree-Cotton and G. Kahane. 2018. The Neuroscience of Moral Judgment. In *The Routledge Handbook of Moral Epistemology*, A Zimmerman, K. Jones, and M. Timmons (Eds.). https://philarchive.org/archive/DEMTNO-9v1.

[17] J.R. Fanchi. 2003. Moral dilemmas: ethics in a senior seminar. In *Proceedings of the 33rd Annual Frontiers in Education (FIE 2003)*. IEEE, S2A1–5.

[18] P. Foot. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5 (1967), 5–15.

[19] M. Frese. 2015. Cultural practices, norms, and values. *Journal of Cross-Cultural Psychology* 46, 10 (2015), 1237–1330.

[20] M. Funk, B. Irrgang, and S. Leuteritz. 2016. Enhanced information warfare and three moral claimsof combat drone responsibility. In *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapon*, E.D. Nucci and F.S.de Sio (Eds.). Routledge, London, UK, 182–196.

[21] B. Gawronski, M. Morrison, C.E. Phills, and S. Galdi. 2017. Temporal stability of implicit and explicit measures: a longitudinal analysis. *Personality and Social Psychology Bulletin* 43 (2017), 300–312.

[22] H.M. Gray, K. Gray, and D.M. Wegner. 2007. Dimensions of Mind Perception. *Science* 315, 5812 (February 2007), 619.

[23] J. Greene. 2014. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them.* Penguin Books, London, UK.

[24] B. Gürçay and J. Baron. 2017. Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning* 23, 1 (2017), 49–80.

[25] C. Harry Hui and H.C. Triandis. 1986. Individualism–Collectivism: A Study of Cross-Cultural Researchers. *Journal of Cross-Cultural Psychology* 17, 2 (1986), 225–248.

[26] E. Hristova and M. Grinberg. 2016. Should Moral Decisions Be Different for Human and Artificial Cognitive Agents. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci2016)*. Cognitive Science Society, Austin, TX, 1511–1516.

[27] J. Ito. 2018. *Why Westerners fear robots and the Japanese do not.* Retrieved November 28, 2020 from https://www.wired.com/story/ideas-joi-ito-robot-overlords/

[28] G Kahane. 2012. On the Wrong Track: Process and Content in Moral Psychology. *Mind & Language* 27, 5 (2012), 519–545.

[29] P.H. Kahn, T. Kanda, H. Ishiguro, T.B. Gill, J.H. Ruckert, S. Shen, H.E. Gary, A.L. Reichert, Freier N.G., and R.L. Sverson. 2011. Do people hold a humanoid robot morally accountable for the harm it causes?. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI2012).* ACM, New York, NY, 33–40.

[30] D. Kasenberg and M. Scheutz. 2018. Norm conflict resolution in stochastic domains. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.*

[31] T. Komatsu. 2016. Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. In *Extended Abstracts of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI2016).* ACM, New York, NY, 457–458.

[32] V. Krishnamoorthy, W. Luo, M. Lewis, and K. Sycara. 2018. A computational framework for integrating task planning and norm aware reasoning for social robots. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* 282–287. https://doi.org/10.1109/ROMAN.2018.8525577

[33] H. Lee, H. Kang, S.S. Kwak, J. Lee, M-G. Kim, and J. Kwon. 2015. How People Perceive Human-and Product-like Robots: Cross-cultural Analysis Between Japan and Korea. In *Proceedings of the 3rd International Conference on Human-Agent Interaction (HAI'15).* 315–318.

[34] J. Lie. 1997. The "Problem" of Foreign Workers in Contemporary Japan. In *he Other Japan: Conflict, Compromise, and Resistance Since 1945*, J. Moore (Ed.). M.E. Sharpe, Armonk, NY, 288–304.

[35] B.F. Malle. 2008. The fundamental tools, and possibly universals, of social cognition. In *Handbook of motivation and cognition across cultures*, R. Sorrentino and S. Yamaguchi (Eds.). Elsevier/Academic Press, New York, 267–296.

[36] B.F. Malle, S. Guglielmo, and A.E. Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (2014), 147–186.

[37] B.F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI2015).* ACM, New York, NY, 117–124.

[38] B.F. Malle, M. Scheutz, J. Voiklis, and J. Forlizzi. 2016. Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI2015).* ACM, New York, NY, 125–132.

[39] B.F. Malle, S. Thapa, and M. Scheutz. 2019. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In *Robotics and Well-Being. Intelligent Systems, Control and Automation: Science and Engineering*, M.I. Aldinhas Ferreira, J. Silva Sequeira, V. Gurvinder, O. Tokhi, and E Kadar (Eds.). Springer, 111–133. https://doi.org/10.1007/978-3-030-12524-0_11.

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

[41] M. Mori. 1989. *The Buddha in the Robot: A Robot Engineer's Thoughts on Science and Religion.* Kosei Publishing Co., Tokyo, Japan.

[42] H. Nitto, D. Taniyama, and H. Inagaki. 2017. Social Acceptance and Impact of Robots and Artificial Intelligence – Findings of Survey in Japan, the U.S. and Germany. *NRI Papers* 211 (February 2017), 1–15.

[43] T. Nomura, T. Suzuki, T. Kanda, J. Han, N. Shin, J. Burke, and K. Kato. 2008. What people assume about humanoid and animal-type robots: cross-cultural analysis between Japan, Korea, and the United States. *International Journal of Humanoid Robotics* 5, 1 (2008), 25–46.

[44] I. Patil, M. Calò, F. Fornasier, F. Cushman, and G. Silani. 2017. The behavioral and neural basis of empathic blame. *Scientific Report* 7, 1 (2017), 5200.

[45] R. Rodogno. 2017. Robots and the Limits of Morality. In *Social Robots – Boundaries, Potential, Challenges*, Marco Nørskov (Ed.). Routledge, London, UK, 17 pages. https://doi.org/10.4324/9781315563084.

[46] Vasanth Sarathy, Matthias Scheutz, and Bertram F. Malle. 2017. Learning behavioral norms in uncertain and changing contexts. In *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom).* 301–306.

[47] H. Sauer. 2012. Morally irrelevant factors: What's left of the dual process-model of moral cognition? *Philosophical Psychology* 25, 6 (2012), 783–811.

[48] M. Scheutz and B.F. Malle. 2014. "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).* 36–39.

[49] M. Scheutz and B.F. Malle. 2020. May Machines Take Lives to Save Lives?: Human Perceptions of Autonomous Robots (with the Capacity to Kill). In *Lethal Autonomous Weapons: Re-Examining the Law & Ethics of Robotic Warfare.*

[50] K.G. Shaver. 1985. *he Attribution of Blame: Causality, Responsibility, and Blameworthiness.* Springer Verlag, New York, NY.

[51] Y. Sone. 2017. *Japanese Robot Culture: Performance, Imagination, and Modernity.* Palgrave Macmillan US, New York, NY.

[52] R. Sparrow. 2007. Killer robots. *Journal of Applied Philosophy* 24, 1 (2007), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

[53] H.C. Triandis, R. Bontempo, M.J. Villareal, M. Asai, and N. Lucca. 1988. Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships.

Journal of Personality and Social Psychology 54, 2 (1988), 323–338.

[54] J. Voiklis, B. Kim, C. Cusimano, and B.F. Malle. 2016. Moral judgments of human vs. robot agents. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN’16)*. 775–780.

[55] Chris Wargo, George Hunter, Ray Young, and Lance Sherry. 2016. UAS as moral agents: Dilemmas and solutions. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. 1–8. https://doi.org/10.1109/DASC.2016.7777982

[56] Henry M Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72, 3 (May 2001), 655–684. https://doi.org/10.1111/1467-8624.00304

[57] M. Yuki and J. Schug. 2020. Psychological consequences of relational mobility. *Current Opinion in Psychology* 32 (2020), 129–132.