Cognition 254 (2025) 105958

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

People's judgments of humans and robots in a classic moral dilemma

Bertram F. Malle [a,*], Matthias Scheutz [b], Corey Cusimano [c], John Voiklis [d], Takanori Komatsu [e], Stuti Thapa [f], Salomi Aladia [g]

[a] Brown University, Providence, RI 02912, USA
[b] Tufts University, Medford, MA 02155, USA
[c] Yale University, New Haven, CT 06520, USA
[d] Knology, Inc., New York, NY 10005, USA
[e] Meiji University, Chiyoda City, Tokyo 101-8301, Japan
[f] University of Tulsa, Tulsa, OK 74104, USA
[g] New York University, New York, NY 10012, USA

ARTICLE INFO

Keywords:
Morality
Moral judgments
Robotics
Robots
AI
Artificial intelligence
Blame
Norms
Empathy
Perspective taking

ABSTRACT

How do ordinary people evaluate robots that make morally significant decisions? Previous work has found both equal and different evaluations, and different ones in either direction. In 13 studies ($N = 7670$), we asked people to evaluate humans and robots that make decisions in norm conflicts (variants of the classic trolley dilemma). We examined several conditions that may influence whether moral evaluations of human and robot agents are the same or different: the type of moral judgment (norms vs. blame); the structure of the dilemma (side effect vs. means-end); salience of particular information (victim, outcome); culture (Japan vs. US); and encouraged empathy. Norms for humans and robots are broadly similar, but blame judgments show a robust asymmetry under one condition: Humans are blamed less than robots specifically for inaction decisions—here, refraining from sacrificing one person for the good of many. This asymmetry may emerge because people appreciate that the human faces an impossible decision and deserves mitigated blame for inaction; when evaluating a robot, such appreciation appears to be lacking. However, our evidence for this explanation is mixed. We discuss alternative explanations and offer methodological guidance for future work into people's moral judgment of robots and humans.

1. Introduction

Morality is an essential characteristic of human communities. As artificial agents begin to enter these communities, they will, no doubt, encounter morally challenging situations and will be expected to act in ways that people consider morally appropriate. Over the past ten years, we and a number of other researchers have studied people's judgments of artificial agents that make moral decisions and have compared them to judgments of humans who make the same decisions (Hristova & Grinberg, 2016; Laakasuo et al., 2023; Shank et al., 2019; Shank & DeSanti, 2018; Stuart & Kneer, 2021; Sundvall et al., 2023). Understanding these judgments is of urgent societal importance. For machines are starting to not only drive cars but fire missiles (Russell, Aguirre, Javorsky, & Tegmark) or deny bail (Morin-Martel, 2023). Even though they currently do not have any moral competence to appreciate the decisions they make, they may in the future. Designers and engineers cannot simply wait for the impending future of moral robots and then ask people to voice their approval or outrage over the machines' morally significant actions; we must gain an understanding *now* of how people will treat moral robots in the near future.

In 2014, we set out to study how people evaluate robots that make morally significant decisions (Malle et al., 2015). We hoped to gain insights into people's responses to these emerging moral actors and perhaps offer warnings and recommendations about a likely future that includes them. We also hoped to gain insights into moral psychology more generally. If people show systematic differences between morally evaluating robots and humans, we might discover features of human moral psychology that respond flexibly to different agents; conversely, if people treat machines morally the same as humans, we might have evidence for features that are less flexible, lie deeper at the core of moral psychology (perhaps comparable to responses to self-propelled movement or humanlike appearance, Zhao & Malle, 2022).

* Corresponding author at: Department of Cognitive and Psychological Sciences, Brown University, 190 Thayer Street, Providence, RI, USA.
E-mail address: bfmalle@brown.edu (B.F. Malle).

https://doi.org/10.1016/j.cognition.2024.105958
Received 14 June 2024; Received in revised form 2 September 2024; Accepted 7 September 2024
Available online 2 October 2024

This article summarizes ten years of our collaborative research on these questions. The article does not offer a definitive conclusion about how and why people treat robots and humans the same in some respects and differently in other respects; but we do offer some systematic patterns of results, methodological recommendations, and a theoretical sketch that may guide new directions of research.

1.1. Investigating moral machines

What would a machine with moral competence look like? In separate work, some of us examined the major elements of moral competence and asked which properties could and should be implemented in robots and other artificial agents (Malle, 2016; Malle & Scheutz, 2017; Scheutz & Malle, 2017). The elements of adult human moral competence start with having norms and a moral vocabulary, which enable the capacities of moral decision making, moral judgment, and moral communication. In our present studies, the robots that elicited moral evaluations from our participants were portrayed to have most of these capacities—in particular, an understanding of norms and moral decision making. In more recent studies (Malle & Phillips, 2023), we added the additional capacity of moral communication.

Does examining moral evaluations of robots' actions even make sense? It does seem to make sense to people. One early study (Kahn Jr. et al., 2012) found that, in a live setting, a majority of people who were interacting with a robot thought of it as morally accountable for a specific transgressive behavior. Survey research also shows that people ascribe to robots the capacity for moral decision making (Malle & Thapa, 2017; Weisman et al., 2017), and a growing literature has demonstrated that people do morally criticize these agents' decisions (Laakasuo et al., 2023; Malle et al., 2015, 2019; Monroe et al., 2014; Stuart & Kneer, 2021; Sundvall et al., 2023). A number of studies also suggest that people morally evaluate even self-driving cars (Awad et al., 2018; Bonnefon et al., 2016; Franklin et al., 2021; Li et al., 2016; Liu & Du, 2022). However, it is sometimes difficult to determine whom or what people are blaming in this case—the car, the designer, the legislation that permitted the vehicle on the street? When given a choice between assigning responsibility to car or designer, people seem to predominantly hold the designer responsible (Li et al., 2016). When people are asked to evaluate disembodied AI, they are also more reluctant to blame those machines directly (Malle et al., 2019), see them as agents (Wilson et al., 2022), or see any "moral" violation at all (Shank & DeSanti, 2018).

Part of the challenge is that people have no clear conception of the kind of "agent" that an autonomous car, a robot, or an AI is. As a result, they must rely on researchers' descriptions, which creates substantial variation in what kinds of agents are evaluated. For example, the more machine-like an agent is described, the more negative people's moral assessments become (Bigman & Gray, 2018); when an artificial agent is described as competent, its advice is trusted more (Hou & Jung, 2021); and when agents are described as having certain cognitive capacities, people are willing to morally evaluate them and sometimes even judge them more positively than humans (Bigman & Gray, 2018; Kneer & Stuart, 2021; Monroe et al., 2014; Young & Monroe, 2019). Thus, a sensible prerequisite for studying how people make judgments about a robot as a potential *moral* agent is for the robot to be a credible *cognitive* agent—one that has choice capacity, knowledge, and intentions (Stuart & Kneer, 2021). The present studies therefore assessed people's moral evaluations of robots that are capable of making decisions, and moral decisions in particular—but we also tracked whether these moral evaluations made sense to people. More on this point shortly.

A complex, perhaps perplexing literature on moral evaluations of artificial agents has emerged over the past 10 years. Some studies found that people blame these agents less than humans for performing the same actions (Furlough et al., 2021; Gall & Stanton, 2024; Stuart & Kneer, 2021); others found they blame them more than humans (Laakasuo et al., 2023; Liu & Du, 2022; Sundvall et al., 2023); and yet others find no difference (Soares et al., 2023). The factors that differentiate these studies are not well understood; authors have pointed to different kinds of violations, different features of agents, different social relations, and more. Our line of work to be presented here will not allow us to settle the impact of all these differentiating factors, but we hope to identify some critical ones that may guide future research.

Because findings are inconsistent and knowledge is limited, we constrained the problem space. We focus here on robots, not on self-driving cars or virtual, disembodied agents. These (fictitious) robots have considerable social and communicative capacities, and our results are unlikely to generalize to simpler machines. We further constrained our investigations to contexts in which people judged robots entangled in *moral conflicts*—in particular, in moral dilemmas modeled after trolley situations (Foot, 1967; Greene et al., 2001; Petrinovich et al., 1993). Such an approach sets limits on generalizability but has two advantages that help us credibly introduce morally competent robots to participants. First, agents who actively consider and compare the two horns of a dilemma show a grasp of the underlying norms that are in conflict, irrespective of which horn they favor. Second, any decision the agent makes in a moral dilemma will be morally significant and in principle morally defensible (or criticizable). Neither decision is entirely an error or a sign of incompetence; one might disagree with it, but it is a moral decision. In choosing trolley dilemmas, we also wanted to take advantage of some concepts and methodologies that have proven useful in previous research, but we had no interest in diagnosing responses in these dilemmas as "deontological" or "utilitarian" (for critiques, see Gawronski et al., 2017; Kahane et al., 2015).

1.2. Our research approach

Researchers have compared humans and robots on numerous kinds of moral judgments, including whether the actions are appropriate, permissible, wrong, blameworthy, and more (Christensen & Gomila, 2012; O'Hara et al., 2010). This variety is a natural result of the fact that humans really do make different kinds of moral judgments (Barbosa & Jiménez-Leal, 2017; Kneer & Machery, 2019; Malle, 2021; Murray et al., 2024). Because findings on one judgment do not necessarily generalize to findings on another judgment we selected three moral judgments for our studies:

(1) Norm judgments (what the agent should do or is permitted to do);
(2) Moral wrongness judgments (whether the agent's decision was morally wrong or not);
(3) Blame judgments (how much blame the agent deserved for making the decision).

Norm judgments have dominated the research on moral dilemmas. For example, Greene et al. (2001) asked about the "appropriate" choice; Mikhail (2011) used "morally permissible"; Paxton, Ungar, and Greene (2012) asked whether the action is "morally acceptable," and many more (see Christensen & Gomila, 2012, Table 1). Norm judgments take primarily a forward-looking perspective—judgments before the agent makes their decision (Malle, 2021). Such judgments are important for deliberation, anticipation, or persuasion. However, many moral judgments are backward-looking—made after the decision or action occurred. Blame judgments are the paradigmatic case of such judgments, and they directly target the agent: We blame somebody for something they did (Malle, 2021; Malle et al., 2014). Wrongness judgments stand in between, as they can take on either perspective—"This is morally wrong, don't do it" or "This was morally wrong. Why did you do it?" Surprisingly, in hundreds of moral dilemma studies with human protagonists, hardly any asked people to evaluate a protagonist *after* deciding one way or another (but see Everett et al., 2016) or probe for blame judgments (but see O'Hara et al., 2010). These limitations have changed since artificial agents have been included in moral dilemmas (Chu & Liu, 2023; Malle et al., 2015; Sundvall et al., 2023)

In this report, we focus on blame judgments and norm judgments. The

Table 1
Norm judgments in Cluster 1 studies.

Study	Norm probe		Human	Robot	
1.1	permissible	% Action	65.4	73.5	z = 0.90, p = .36
		N	78	49	
1.2	should	% Action	79.3	84.4	z = 1.16, p = .25
		N	184	135	
1.3	should + follow-up	% Action	70.5	83.7	z = 2.9, p = .004
		N	200	196	
1.4	should + follow-up	% Action	78.9	82.6	z = 0.79, p > .5
		N	147	149	
	Cluster 1 total	% Action	74.5	82.6	
		N	609	529	

Note: The percentages show participants favoring Action out of valid participants, excluding those who disqualified the robot from being an independent target of blame (results for total samples are very similar). In Study 1.3 (unlike the other studies), the norm probes were presented after the blame judgment (thus being likely influenced by the human-robot blame asymmetry).

results of wrongness judgments largely parallel those of blame judgments, but the effect sizes are somewhat weaker, in part because fewer than 25% of people considered either decision morally wrong (for more details, see the Supplementary Document, SD). Philosopher Williston (2006) argued that agents in moral dilemmas perform wrong actions but should not be blamed. For ordinary people, the opposite seems to be true.

Asking to make norm judgments is meaningful only when the norms actually apply to an agent; and we assumed that people would naturally consider whether it is *permissible* for a robot to act one way or another, or whether the robot *should* decide one way or another. We tested and verified in Study 1 and later in a study in Japan (Komatsu et al., 2021) that at least 90% of people engage in these considerations. Blaming an agent, however, is meaningful only when the agent is actually a proper *target of blame*—what philosophers have called "having moral responsibility" or "moral agency" (Korsgaard, 2008; Sullins, 2006; Watson, 1982). Some scholars have denied that blame for artificial agents is an appropriate judgment (Sharkey, 2017), but the question here is whether *people* hold a robot morally accountable for its actions, and the initial evidence suggests they do (Banks, 2019; Kahn Jr. et al., 2012; Monroe et al., 2014). However, we wanted to verify this presupposition and therefore included a measure of people's willingness to treat a robot as a proper target of blame in all studies reported here (and also in Malle et al., 2019; Malle & Phillips, 2023).

2. Methods common to all studies

We conducted 13 online experiments (N = 7670 participants). To avoid unwieldy traditional descriptions of individual studies, we group studies together into meaningful clusters. We first summarize the nature of these clusters, then report common methodological features among all studies, and highlight distinguishing features within the specific cluster sections. In a Supplementary Document (SD) we provide further details on methodology, samples, demographics, and additional results. All data are available at https://osf.io/3st2h/.

2.1. Overview of study clusters

Cluster 1 studies introduce the primary finding across all our studies: that people blame humans less than robots when they decide to *not* intervene in a trolley-like moral dilemma ("Inaction asymmetry"). By contrast, we find that people impose very similar *norms* (what is permissible or prescribed) on humans and robots.

Cluster 2 studies examine several boundary conditions to the Inaction asymmetry, including event structure (side-effect vs. means-end), outcome salience, and victim salience.

Cluster 3 studies replicate the Inaction asymmetry in Japan, while also testing what norms Japanese respondents extend to a robot.

Cluster 4 studies examine the hypothesis that the Inaction asymmetry may be best explained by a kind of empathic mitigation of blame for human agents not extended to robot agents.

2.1.1. Participants

We recruited participants from online crowdsourcing platforms, such as Amazon Mechanical Turk, Prolific, and Yahoo! Japan, as well as one student sample. Details of each sample can be found in the Supplementary Materials document.

2.1.2. Procedures

In all studies, participants received at most a brief introduction (e.g., "On the next page you will read a short story…"). Then they read the main narrative, which was presented one paragraph at a time. After the dilemma was set up, people were asked two moral judgments. In six studies, a *norm judgment* (e.g., Is the action permissible? or What should the agent do?) preceded a description of the agent's *decision*, which was followed by a *blame* judgment. In the remaining studies, participants learned about the decision and then made both a wrongness and a blame judgment.

The experimental conditions of Agent (human or robot) and agent's Decision (action or inaction) were manipulated between subjects. After providing their moral judgments, participants were asked to explain one or more of their judgments, and they always explained blame judgments. For these blame judgments (made on a 0–100 scale), the prompt for explanations was "Why do you think the [robot | repairman] deserves this amount of blame?" At the end, we collected demographics and various exploratory measures (detailed in the SD).

2.1.3. Materials

We modeled our studies after the "trolley dilemma" paradigm (Christensen et al., 2014; Foot, 1967; Petrinovich et al., 1993) but modified it somewhat to easily set up a robot's involvement. The basic narrative is as follows (variations between studies are culled in the SD):

A runaway train with four workers on board is about to crash into a wall, which would kill all four, unless the protagonist (a repairman or repair robot) performs an action (e.g., redirecting the train in most studies or dropping a cart onto the tracks in three studies) that saves the four. As a result of the action, however, a single worker would be killed. Participants thus evaluate a protagonist who (i) decides to take a specific action that saves four people but causes a single person to die ("Action") or (ii) decides to not take that action, spare the one person, but allow the four to die ("Inaction"). We always used the word *decide* because we wanted to highlight the intentionality of *either* path, rather than create a full-blown action-omission case.[1]

In the wording of the narrative, we described the protagonist with at least two mental state verbs (*spot*, *recognize*) as well as the verb *decide*, because we assumed that a robot with credible *cognitive* capacities would be a candidate for having credible *moral* capacities (Bigman et al., 2019; Monroe et al., 2014; Stuart & Kneer, 2021).

2.1.4. Data treatment and statistical analysis

Identifying people who disqualify robots as targets of blame. Already in our first study, we discovered participants who expressed that they disqualified the robot agent as a proper target of blame. In their explanations of blame judgments, they spontaneously mentioned that a robot "doesn't have a moral compass," "cannot make moral decisions," "is not a person," "is merely programmed." Indeed, about a third of participants

[1] The structure of trolley-like dilemmas is problematic if one wants to draw conclusions about deontological vs. utilitarian tendencies (Gawronski et al., 2017). We had no interest in such conclusions; we were interested in comparing a robot's and a human's decision to act in one way or another, whereby both paths are morally significant because they invoke and violate moral norms. We will return to these issues in the General Discussion section, addressing limitations and future research directions.

disqualified the robot in this way. Averaging blame ratings from those who do and those who do not find blame for robots meaningful distorts the results. In particular, those who deny robot moral agency predominantly provide 0 or low ratings, which, when averaged with valid ratings, can give the illusion that robots are blamed less. For all these reasons, we adopted a systematic coding process of identifying such disqualifying statements and applied it to the present clusters of studies (and also in Komatsu et al., 2021; Malle et al., 2016, 2019). See SD for details. All coded responses are available at https://osf.io/3st2h/.

Hypothesis tests for blame. Our original approach was to test the hypothesis of an interaction between Decision (action-inaction) and Agent (human-robot) (Malle et al., 2015), but we occasionally also reported simple effects (e.g., Malle et al., 2016). Increasingly, the patterns of findings convinced us to focus on a pair of simple effects — a possible human-robot asymmetry for inaction decisions and a possible asymmetry for action decisions —while also documenting the interaction for completeness. We report here significance tests for the two simple-effects hypotheses as well Cohen's *d* effect sizes for the two hypotheses and for the interaction. (See SD for a detailed explanation of computing Cohen's *d* for interaction terms).

3. Cluster 1 studies: human-robot asymmetries

3.1. Goals and main features of studies

The data composing Study 1.1 were initially published in Malle et al. (2015), but we are reporting them here with a few changes detailed in the SD. The study documented, for the first time, that people might impose similar norms on human and robot agents but blame humans less for inaction (not intervening in the dilemma) and, potentially, blame robots less for action.

In Studies 1.2 to 1.4, we attempted to replicate the asymmetry of blame judgments and examined more deeply the pattern of norm judgments. Specifically, Study 1.2 replaced the frequently used permissibility question with the question, "What should the [repairman]/[robot] do in this situation?" Studies 1.3 and 1.4 also asked participants to further clarify what they meant by their response to the "should" question. We offered several previously validated expressions from Malle (2020), and people could choose which of them best fit their initial assessment. The expressions included permission terms (*acceptable, permitted, optional)* and prescription terms of increasing strength (*called for, essential, required, mandatory)*. For more details, see the SD.

3.2. Results of Cluster 1

3.2.1. Norm judgments

Many people (79.7% overall) endorsed the decision to switch the train and save four people (see Table 1). Around this mean, we found a small human-robot difference such that more people preferred for robots to make the switch (82.6%) than for humans to do so (74.5%). This difference was consistent but significant in only one study, namely when norm judgments followed blame judgments (Study 1.3). Apparently, because people blame humans and robots differently (see below), these blame judgments pulled norm judgments into the same direction.

The more differentiated assessment of norm judgments in Studies 1.3 and 1.4 is described in detail in the SD. In summary, it showed that when people indicate that an agent "should" make a decision in this dilemma, three fourths of them mean something weaker: that it is *permissible* to so act. This tendency toward endorsing a permission rather than a prescription was somewhat greater for the human agent, but again only in Study 1.3, when norm judgments followed blame judgments. When we examined those participants who indicated a prescription rather than merely a permission, we found no human-robot differences in the strength of those prescriptions in either 1.3 or 1.4. All in all, we see that norms people impose on robots are surprisingly similar to those they impose on humans (at least in this kind of moral dilemma). Evidence that people prefer robots to act (i.e., switching the train and sacrificing a single individual) is weak and magnifies only under the influence of prior blame judgments.

3.2.2. Blame judgments

A blame asymmetry emerged consistently, which can be captured by an interaction term (which is significant in Studies 1.1, 1.2, and 1.3; see Table SD7 for all means and significance tests) but is clearly a two-fold pattern, as Fig. 1 demonstrates: When the agent decides to *not* act, people blame humans less for this inaction decision than they blame robots (Cohen's *d* value for this difference range from 0.44 to 0.70); when the agent decides to *act*, there is no human-robot difference. Mean blame ratings are stable at just over 40 (on a 0–100 scale) in three of the conditions—human action, robot action, and robot inaction—but are 15 points lower in the condition in which the human chooses inaction. We may therefore consider this a *mitigation* effect—reduced blame for a human deciding not to make the tough choice of sacrificing one person to save four.

3.3. Discussion of Cluster 1 results

We take three insights away from this first cluster of studies. First, humans and robots differ only minimally in the kinds of norms people impose on them (consistent with Malle et al., 2019), but they do differ in how much blame people assign to them. Blame for a human agent is somehow mitigated when the person decides to *not* intervene in the dilemma. We will examine in detail what might explain this pattern after we explore boundary conditions (Cluster 2) and its possible generalization beyond U.S. culture (Cluster 3).

A second insight concerns moral judgments more generally, namely, that single norm judgments (e.g., permissible, should) can be misleading. When participants were forced to select which of two paths in the dilemma an agent should take (prescription), 75%–79% recommended Action. When we asked them to choose from a wider array of options, including terms of permission and various degrees of prescription (see SD for details), three fourths of these participants moderated their judgment and declared that the chosen path is only *permissible*. When we subsequently inquired about the alternative path, using an array of permission and prohibition terms, more than half of participants expressed that this alternative was also permissible (even though they had rejected it in response to the should question). Thus, when participants declare that, say, the *action* path in a dilemma is prescribed or permissible we cannot conclude that they find the alternative—inaction—*impermissible*. Drawing conclusions about deontological and utilitarian attitudes from such judgments would seem to be tenuous.

Finally, the small human-robot difference for norms and the larger and robust one for blame judgments provides further evidence for the distinct nature of norm and blame judgments. In hindsight, this may not be all that surprising but has not been fully appreciated in the moral psychology literature (Malle, 2021). It was especially overlooked in the study of moral dilemmas, where blame judgments were almost never probed. We might gain novel insights into both moral dilemmas and moral judgment if we distinguish norm from blame judgments.

4. Cluster 2 studies: boundary conditions

4.1. Goals and main features of studies

In this cluster of four studies, we examined a number of possible boundary conditions to the human-robot asymmetry for blame judgments found in Cluster 1. First, we tested the classic distinction between a side-effect scenario and a means-end scenario of the trolley problem (Feltz & May, 2017; Greene et al., 2009; Levine et al., 2018; Mikhail, 2009). We call this comparison *event structure.* In a side-effect structure (which we had employed in Cluster 1 studies), the death of one person is

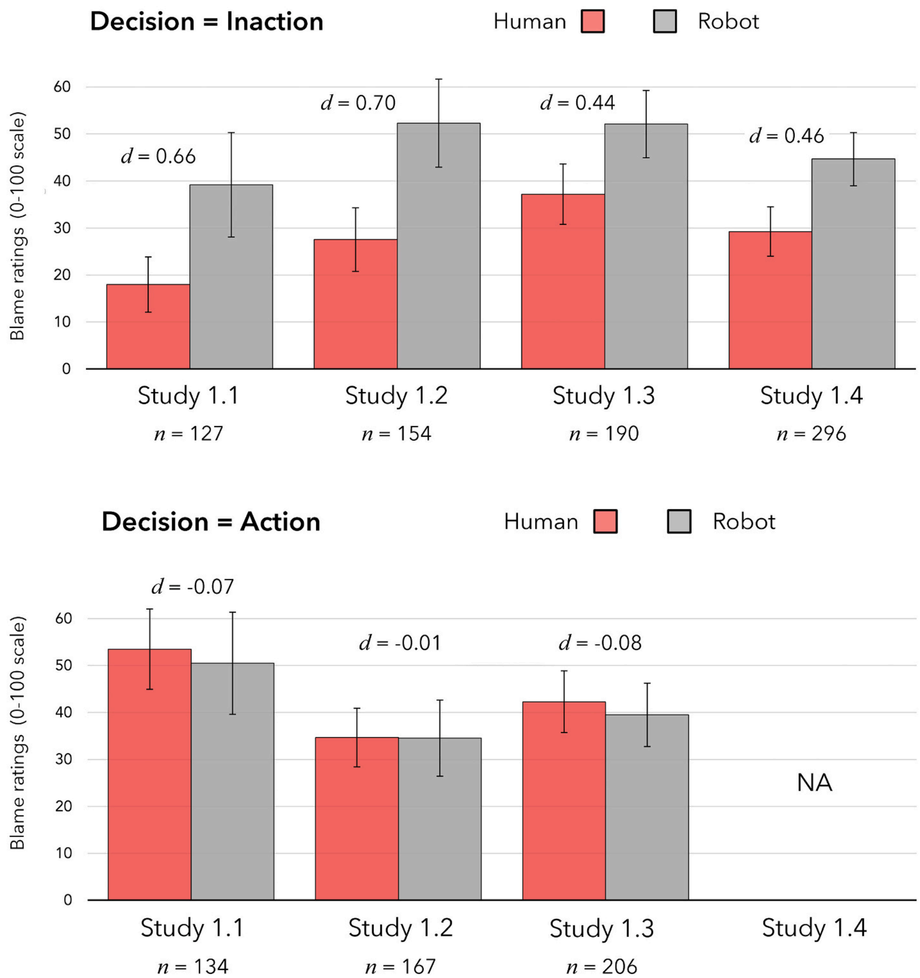


Fig. 1. Means (and 95% CIs) of blame ratings in Cluster 1 studies. Upper panel: Tests of the Inaction asymmetry (where humans are blamed less than robots for deciding to *not* act). Lower panel: Tests for a possible Action asymmetry. Indicated sample sizes are those on which tests are based (not counting the participants in the other decision condition).

an inevitable side effect of one's attempt to save the four. In a means-end structure, the agent directly uses the one worker as a means to the end of saving the four (see Table 2).

Second, we examined the salience of lives saved and lost, which we call *outcome salience*. Wrongness and blame judgments are independently sensitive to variations of mental states (e.g., beliefs, decisions) and variations of outcomes (e.g., one person vs. four people dying) (Cushman, 2008; Young & Saxe, 2009). In our studies we did not vary the *severity* of outcome (e.g., 4 vs. 10 people dying) but rather the *salience* of the casualties—by mentioning who and how many died or leaving that outcome implicit (see Table 2).

Third, we examined a factor we call *victim salience*. We had noticed in our initial studies that the phrasing of the focal action (what the agent decides to do, or not to do) might have an impact on the results: Action phrasing that highlighted the victim as a target ("direct the train toward the single miner") sometimes weakened the human-robot asymmetry compared to phrasing that did not mention the victim ("switch the train onto the side rail").

Fourth, in Study 2.4 we made one change that we hoped would not be a boundary condition but allow generalization. We devised a dilemma in which the action path is more objectionable (even without outcome or victim salience; we refer to this dilemma as the "chute" scenario). In the scene, the initial setup is the same as before, but the necessary action to slow down the train is to open a chute that either drops a cart onto the track (along with, inevitably, a worker)—which is the side-effect structure—or drops the worker himself onto the track—which is the means-end structure. As expected, participants see this scenario as especially challenging: about half of them recommend action while the other half recommend inaction (see norm judgment results in Cluster 3).

In total, we conducted four studies in this cluster, which started with an exploration and became increasingly more systematic. Study 2.1 ($N = 159$ after exclusions) was the initial exploratory study, where (we now know) means-end structure, outcome salience, and victim salience co-occurred. Study 2.2 ($N = 456$), maintained outcome salience and experimentally manipulated the two types of event structure. Study 2.3 ($N = 774$) had no outcome salience and experimentally manipulated both event structure and victim salience. Study 2.4 ($N = 640$) had neither outcome nor victim salience, and we experimentally manipulated event structure, this time in a variant of the original dilemma in which the action was no more preferred than the inaction.

We report here the results in outline, and all means, effect sizes, and statistical tests are available in Table SD10.

4.2. *Results of Cluster 2*

4.2.1. *Event structure*

We tested this contrast between means-end and side-effect structure in Study 2.2 (crossed with outcome salience), in Study 2.3 (crossed with victim salience), and in Study 2.4 (without outcome and victim salience, and in a somewhat different dilemma). The results show a consistent pattern: When randomly assigned, side-effect scenarios show larger

Table 2
Sample text from manipulation of potential boundary conditions to human-robot blame asymmetry in Cluster 2 studies: Event structure, outcome salience, and victim salience.

1. Event Structure	
Side Effect	Means-End
The [repairman | robot] recognizes that if the train continues on its path it will crash into a massive mine wall and kill the four miners. If it is switched onto a side rail, it will kill a single miner who is working there while wearing headsets to protect against a noisy power tool.	The [repairman | robot] also recognizes that the four miners can be saved if something slowed down the train. In fact, if the train were directed onto a side rail, it would strike a single miner who is working there, wearing headsets to protect against a noisy power tool. The train would hit and kill the single miner, it would slow down as a result, and the four miners on the train would survive.
2. Outcome Salience	
Not Salient	Salient (lives saved and lost)
[Action:] The [repairman | robot] decides to direct the train onto the side rail.	[Action:] The [repairman | robot] decides to direct the train onto the side rail. The train strikes and kills the single miner; the four miners on the train survive.
3. Victim Salience	
Not Salient	Salient (victim as a target)
In fact, the [repairman | robot] decided to [not] switch the train onto the side rail.	In fact, the [repairman | robot] decided to [not] direct the train toward the single miner.

inaction effects (Cohen's d) than means-end scenarios: 0.43 > 0.34, 0.69 > 0.35, 0.37 > −0.14, 0.38 > 0.27.

4.2.2. Outcome salience

Scenarios with explicit outcome information (about who and how many survived or died) appeared in Study 2.1 and in two conditions of Study 2.2. In Study 2.1, it co-occurred with both means-end structure and victim salience, and that joint impact reversed the means pattern (Inaction asymmetry $d = -0.26$), though the effect did not significantly go in the opposite direction. In Study 2.2, outcome salience co-occurred with means-end structure in one condition ($d = 0.34$) and with side-effect structure in the other condition ($d = 0.43$), and these effect sizes are within the range of several of our other studies. It therefore appears that outcome salience is at most mildly detrimental.

4.2.3. Victim salience

Exploratory Study 2.1 included both victim salience, means-end structure, and outcome salience, and jointly these three conditions pushed the Inaction asymmetry toward reversal ($d = -0.26$). We manipulated victim salience crossed with event structure systematically in Study 2.3, in a 2 (Event structure) × 2 (Victim salience) × 2 (Agent) × 2 (Decision) design. The Inaction asymmetry was strongest and significant in the condition featuring a side-effect structure without victim salience, and the full Agent × Decision interaction was also visible and significant only in this condition. The other three conditions (means-end structure with or without victim salience) eliminated any human-robot asymmetry (see Fig. 2).

Another way to understand the systematic patterns in Study 2.3 is by displaying the patterns of effect sizes for inaction, action, and their statistical interaction (see Table 3). Under a side-effect structure and when the victim is not salient, the effect sizes are as high as in Cluster 1. When either the means-end structure or victim salience enter the scenario, the inaction effect drops and the full interaction disappears. And when both means-end structure and victim salience co-occur, the effect practically reverses. In addition, we see that the *Action* asymmetry begins to grow with the boundary conditions present — that is, the robot is blamed increasingly (and more than the human agent) when the action has a means-end structure and/or is victim-directed.

4.2.4. All boundary conditions viewed jointly

Fig. 3 summarizes how the boundary conditions across all studies in Cluster 2 affect the Inaction asymmetry, but broken down not by Studies

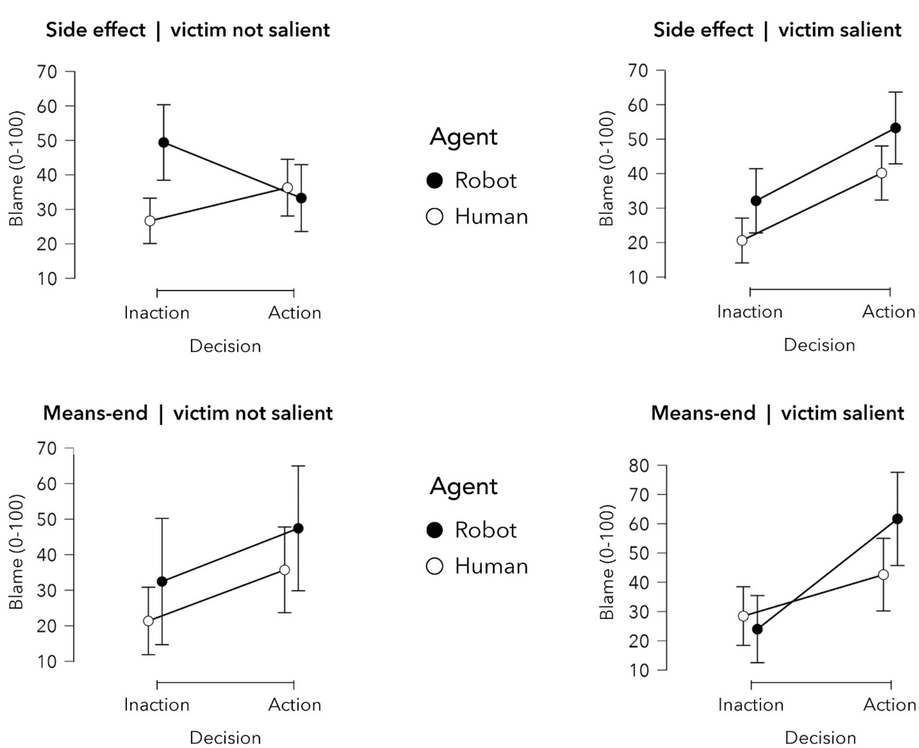


Fig. 2. Means and 95% CIs for blame in four conditions of Study 2.3. The Inaction asymmetry is strong and significant only in the left upper condition (side-effect structure, no victim salience). The remaining three conditions show no inaction asymmetry.

Table 3
Cohen's *d* effect sizes in Study 2.3 as a function of its two crossed boundary conditions, Event structure and Victim Salience.

	Inaction *d*		Interaction *d*		Action *d*	
	Event Structure		Event Structure		Event Structure	
Victim Salient	Side effect	Means-End	Side effect	Means-End	Side effect	Means-End
No	0.37	−0.14	-0.20	−0.34	0.36	0.52
Yes	0.69	0.35	0.37	−0.01	−0.08	0.30

but by conditions that contain one or more of the boundary conditions. As the number of co-occurring boundary conditions decreases from three to two to one, the inaction effect begins to turn in the predicted direction (robot > human) and becomes consistently significant under side-effect structure without any salience manipulations.

4.3. Discussion of Cluster 2

The pattern of moral judgments in Cluster 1 showed minimal human-robot difference for norm judgments (permissibility, should) but we found a robust Inaction asymmetry for blame judgments. Cluster 2 identified systematic boundary conditions to this asymmetry, which hold when the action is highly objectionable because the protagonist uses a person as a means to an end and/or targets the victim. Under these conditions, people increase blame for any protagonist who undertakes such instrumental harm, but they particularly object to a robot doing so. Thus, people blame the robot increasingly for action and less for the justifiable response of inaction; as a result, the Inaction asymmetry weakens or disappears. This pattern is consistent with Laakasuo et al.'s (2023) finding that robots receive particularly strong disapproval for violating human autonomy (e.g., by following orders to forcefully medicate a patient).

By contrast, in all of Cluster 1 studies and the side-effect scenarios in Cluster 2, action is generally favored and though people on average still blame agents for it, robots and humans are blamed the same amount. For the choice of inaction, however, human agents get a pass: People mitigate their blame for the human who "cannot" decide. This mitigation may be the result of empathy with the person's terrible decision conflict (Gamez-Djokic & Molden, 2016; Rom et al., 2017), and we will take up the possibility of this empathy-based explanation in Cluster 4.

First, however, we report on a cluster of studies that sought to explore generalization of the effect. We examined whether the similarity in norms and the Inaction asymmetry for blame judgments would replicate in a culture distinct from the U.S. We chose Japan for its technological advances that make evaluations of moral robots plausible, for its dissimilarity from the U.S. on known cultural dimensions (Gelfand et al., 2006; Triandis et al., 1988), and because one previous study in Japan (Komatsu, 2016) had shown different results from our original finding in Malle et al. (2015).

5. Cluster 3 studies: culture

5.1. Goals and study features

The three studies in this cluster (3.1 to 3.3) were originally reported in Komatsu et al. (2021), comparing Japanese and U.S. participants in the two dilemma scenarios we have studied here: the chute dilemma (in Study 2.4) and the standard switch dilemma (used in all other studies). We summarize the motivation and main results in light of the previous clusters' and focus on three points.

Primarily, the cross-cultural project asked whether the Inaction asymmetry for blame judgments replicates in an East Asian sample. Our working hypothesis at the time considered the Inaction asymmetry a result of dampened social-cognitive inferences toward robots (Malle et al., 2019; Scheutz & Malle, 2021). Because there was no a priori reason to expect such inferential activity to differ between cultures we expected the Inaction asymmetry to hold in both Japan and the U.S.

We have argued that a critical requirement for testing any human-robot blame asymmetry is to identify, and exclude from analysis, those participants who spontaneously declare that a robot is not a proper target of blame. We therefore examined whether the rate of those participants is comparable in the two cultures, and we speculated that Japanese participants would show a lower rate of disqualification because of the greater acceptance of robots in Japan (Sone, 2017). However, once correcting for disqualifications, the Inaction asymmetry should still hold.

In addition to blame judgments, we tested whether the *norms* (measured as permissibility judgments) for intervening in the dilemmas

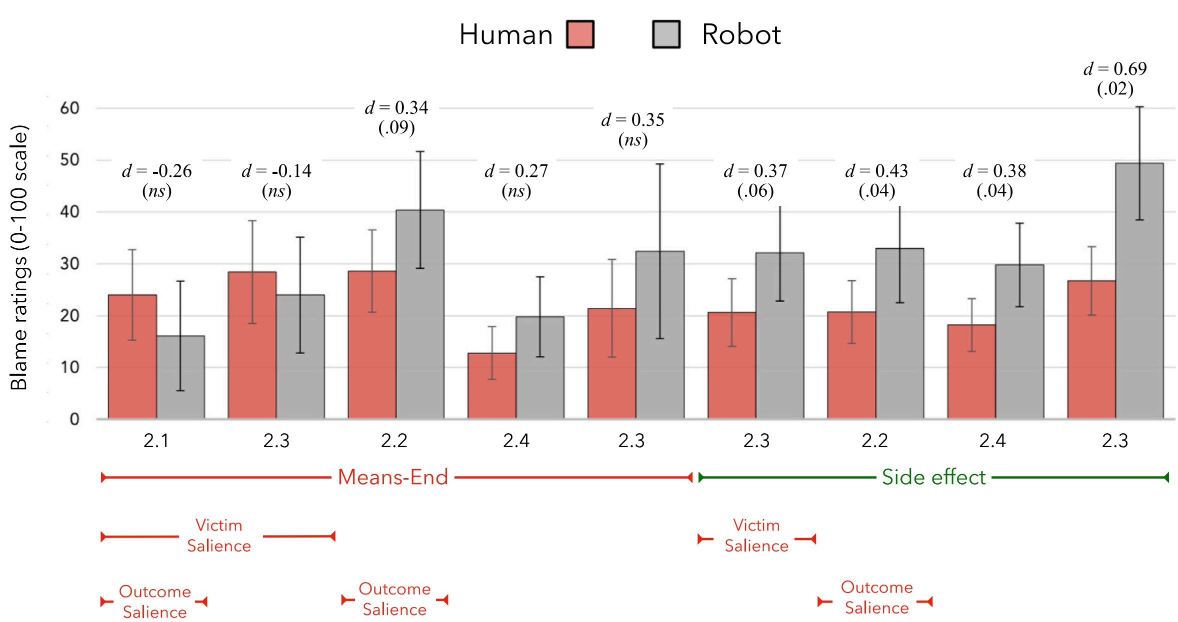


Fig. 3. The Inaction asymmetry across all Cluster 2 studies (2.1 to 2.4) and their conditions. As the number of boundary factors decreases, the effect begins to turn consistently in the predicted direction (robot > human) and turns significant under side-effect structure.

are similar in the two cultures. Many norms differ between the cultures (Nitto et al., 2017; Triandis et al., 1988), but which ones apply to moral dilemmas of the tested kind was less obvious. Perhaps Japanese participants more strongly favor the collective (i.e., they would support the decision to protect four rather than one) or disfavor an individual's autonomous intervention in the tragic process (i.e., they would support the decision to *not* intervene). Either way, because norms seemed largely independent of blame in our previous studies, we expected that the Inaction asymmetry for blame would hold whether or not any norm differences emerged.

We examined these hypotheses in three samples (for details on Methods, see Komatsu et al., 2021). First, we replicated the standard switch dilemma in a Japanese sample (Study 3.1) and compared the results to an aggregate of previously collected U.S. samples. Then we examined Japanese norm and blame judgments for the chute dilemma (Study 3.2) and simultaneously conducted the study in a U.S. sample (Study 3.3). The latter study was similar to Study 2.4's side-effect condition but had minor phrasing changes and included a norm question.

5.2. Results of Cluster 3

We highlight important results below and offer more detailed results in Tables SD13 to SD15. For consistency, we adopt here the same analysis approach as in all other studies in this article, so some numbers will slightly deviate from the Komatsu et al. report.

First, we found lower disqualification rates in the Japanese samples (15.4 and 16.9%) than we had seen in all our previous studies (32.5%) as well as in the new U.S. study (25.1%). We infer that Japanese participants are not only more accepting of robots generally but apparently also of robots that are cast into moral decision making roles.

Second, we found modest variations in norm judgments. Both cultures saw the action decision in the chute dilemma about 14% less morally permissible than the action decision in the switch dilemma. Japanese participants considered the active intervention in both dilemmas about 10 percentage points less morally permissible than U.S. participants. In both cultures, small human-robot differences (near 10 percentage points) emerged in the switch dilemma, but only Japanese participants continued to display a similarly sized difference in the chute dilemma.

Third, the Inaction asymmetry for blame replicated in the Japanese samples. In the switch dilemma, the effect size was smaller ($d = 0.29$) than we had seen in our aggregated studies ($d = 0.45$). The chute dilemma effect size in Japan was similar to Study 2.4, but the newer U.S. sample showed a weaker effect size ($d = 0.24$) in this particular sample.

The somewhat smaller effect sizes in these studies becomes larger when we test the asymmetry among those who declared the action decision in the dilemma to be permissible. For these participants, both action and inaction seem to be viable paths, and the average blame for both options is below 50 (on the 0–100 scale). By contrast, those who consider the action to be *impermissible* seem to be expressing a prohibition and therefore give very high blame ratings for the action decision (because it violates the prohibition) and very low blame ratings for the inaction decision (because it is the only option that does not violate the prohibition). With these very low blame ratings, the chance to detect a human-robot difference runs up against a floor effect. (In Study 1.1, the other sample in which permissibility was measured, the inaction asymmetry was also weaker among those who considered the action impermissible.)

5.3. Discussion of Cluster 3

Our results suggest that, for the moral dilemmas we examined, Japanese and U.S. participants differed in two ways: Fewer Japanese participants disqualified the robot as a proper target of blame, and fewer Japanese participants considered action in the two dilemmas to be permissible. However, they responded similarly in several other ways: They found the chute dilemma less permissible than the switch dilemma; they found action in those dilemmas to be slightly more permissible for the robot than for the human; and they consistently blamed the human less than the robot for the inaction decision (though effect sizes were smaller than in previous studies). Thus, the Inaction asymmetry generalizes to at least one non-U.S., collectivist culture.

6. What explains the Inaction asymmetry?

We have seen in over ten studies that the Inaction asymmetry for blame is a robust phenomenon: Except when the action under consideration is highly objectionable (e.g., because a person is used as a means to an end), people blame humans less than robots when they refrain from acting in a classic dilemma—that is, refrain from sacrificing one for the good of many. What explains this asymmetry?

A first possibility is that people do not blame humans less; they blame robots more. They expect robots, more so than humans, to act as "utilitarians" and to save the largest number of lives (Zhang et al., 2022); when the robots don't act as utilitarians, they get blamed more than humans. (This is the original interpretation we had adopted in 2015.) This account is supported to some extent by the pattern of norm judgments, where people have a slight preference for robots to choose action. However, this preference is weak in absolute terms (averaging around 6% across all studies we assessed). What further speaks against the hypothesis is that the pattern of means suggests humans are being blamed less, rather than robots being blamed more, for inaction decisions. In the studies reported so far, when the Inaction asymmetry effect size was at least above zero, the average blame for action was 41.9 for the human and 43.9 for the robot; the average blame for inaction was 35.2 for the robot and 21.9 for the human—the latter being by far the lowest of the four numbers. The third and final reason to doubt the robot-utilitarian account is that, if robots are envisioned more as utilitarians than humans are, then a robot that chooses *action* (which is in line with the utilitarian ideal) should receive less blame than a human who does so. In reality, blame for human and robot agents was consistently similar in this condition. In scenarios with means-end event structure or victim salience, the robot was even blamed more than the human. Sundvall et al. (2023) also cast doubt on the utilitarian analysis. They assessed moral judgments of robots and humans who had to choose whom to save from an accident at sea—one person vs. two people, and ones culpable for the accident vs. innocent victims. The utilitarian consideration of how many lives were saved influenced moral approval of both robots and humans, whereas the nonutilitarian consideration of culpability of the person(s) saved was consistently more important for robots.

An alternative account of the Inaction asymmetry is this. When people blame agents for intentional behaviors (such as the decisions in the present dilemmas) they infer the agent's reasons and motives (Carlson et al., 2022; Cushman, 2008; Malle et al., 2014). So when people blame an agent less, they may have inferred more charitable reasons—reasons that help justify the person's decision (Scheutz & Malle, 2021). What might such charitable reasons be?

To explore potential reasons that people ascribe to the protagonists' decisions we inspected people's explanations following their blame judgments across the studies reported so far (for details on the coding method, see SD). Two frequently mentioned groups of words emerged: one referred to intentionality and choice; the other referred to the difficulty of the decision. The intentionality group occurred more frequently but did not differentiate between agents in the inaction condition ($\chi^2 < 1$), where the blame asymmetry of interest exists. But explanations referring to the decision's difficulty showed a strong pattern in the Inaction condition: those participants who judged a human spontaneously mentioned the decision's difficulty almost twice as often (15.3%) as participants who judged a robot (8.0%). They highlighted the "impossible decision, "terrible choice", "horrible situation, or "tragedy." Thus, one interpretation of the blame mitigation for human inaction decisions is that people empathize with the human protagonist's agony of the choice dilemma, understand his inaction decision, and therefore find it defensible. This third-person process is consistent with a finding in Gamez-Djokic and Molden (2016), where first-person reported difficulty with similar moral dilemmas predicted a preference for inaction choices. It is further consistent with Rom et al. (2017), who found that people ascribe more affective than cognitive processes to a person who makes an inaction decision and also ascribe more warmth and morality to that person. The much higher rate of mentioned "difficulty" for human than robot protagonists in our studies might also reflect people's ability to recognize and appreciate constraints on other people's reasoning (Cusimano, Zorrilla, Danks, & Lombrozo, 2024; Cusimano & Goodwin, 2020) and a resulting inclination to ascribe more favorable dispositions to them. These processes are less likely to emerge when encountering robot protagonists whose reasoning people do not understand and to whom they therefore do not extend the kind of mitigation they extend to humans. We call this the "empathy hypothesis" but consider its label a convenient shortcut rather than a postulate of a specific process.

With these considerations in mind, in Cluster 4 we attempted to induce people to consider the robot's difficult choice and, in understanding the challenge of its decision, to conjure up charitable reasons for the robot's inaction.

7. Cluster 4 studies: the empathy hypothesis

7.1. Study 4.1

This study was the first attempt to examine whether we could induce people into empathizing with the "plight" of the robot, perhaps mitigating their blame for its inaction decision. We used the aggregate means of Cluster 1 studies to provide the comparison standard for this empathy manipulation. We exposed 575 participants (after exclusions) to a side-effect scenario in which the last paragraph was replaced with this text: "Having to decide whether or not to switch the train onto the side rail, the [repairman | robot] struggles with the difficult decision. But time is running short; the [repairman | robot] needs to make a choice." For an additional 208 participants, we replaced the phrase "struggles with" with "deliberates about," as an exploratory condition that made the mind of the robot salient without referring to an emotional state.

The "struggle" and "deliberate" conditions showed identical effect sizes of $d = 0.25$, with an overall Inaction asymmetry of $d = 0.25$, $F(1,635) = 4.6$, $p = .032$. This asymmetry is about half the size of the asymmetry in the aggregate of Cluster 1 ($d = 0.54$). A test of Study 4.1's Inaction asymmetry against the aggregate asymmetry in Cluster 1 was significant, $F(1, 1910) = 23.4$, $p < .001$. Moreover, the shift in means occurred specifically in the robot condition. Whereas the average blame for robots choosing inaction in Cluster 1 was 47.2, the average in Study 4.1 was reduced to 36.0; the human means barely changed, from 28.7 in Cluster 1 to 27.4 in Study 4.1.

7.2. Study 4.2

We then designed and preregistered a highly powered second study (https://osf.io/dqr54), attempting to replicate the struggle manipulation and randomly assigning participants to either this manipulation or a standard side-effect condition as a control. We limited ourselves to the important inaction decision (where the manipulation is expected to operate). We slightly rephrased the struggle manipulation: "Deliberating whether or not to switch the train onto the side rail, the [repairman | robot] struggles with the extremely difficult decision." We also included a norm question (what the agent *should* do), which yielded a preference for the robot to choose action (86.1%) compared to a human to choose action (75.8%), $z = 2.37$, $p = .009$.

In addition, we introduced six new rating items to probe people's (a) reported engagement in active mental simulation when reading the scenario, (b) perceptions of the difficulty of the choice, and (c) understanding of the agent's decision (2 items each; see https://osf.io/35w9c and SD for details). We preregistered analyses to examine whether these perceptions might mediate the effect of agent type and condition on blame.

Blame judgments in the control condition replicated the familiar human-robot Inaction asymmetry at $d = 0.35$, $F(1,230) = 5.7$, $p = .018$. However, the struggle manipulation did not change blame for the robot and yielded the same human-robot asymmetry at $d = 0.37$ ($p = .016$).

Earlier we had introduced the content-coded variable of *Mentioned difficulty*—the frequency of people mentioning the difficulty of the "impossible" decision. We used it here as a manipulation check, examining whether the struggle manipulation increased the rate of mentioned difficulty. Indeed, collapsed across agent type, the frequency increased from 9.9% (in the control condition) to 17.8% (in the struggle condition), $z = 2.17$, $p = .03$. And collapsed across control and struggle conditions, this frequency was considerably higher for human agents (19.4%) than for robot agents (11.8%), $z = 2.91$, $p = .004$. We also expected an interaction effect, such that the struggle manipulation would particularly affect people in the robot condition, which was designed for them to "catch up" in their empathy with the robot. But there was an opposite trend: While people judging a robot mentioned its difficulty more often in the struggle condition (14.0%) than in the control condition (9.6%), this difference was even stronger for humans (28.8% vs. 10.8%), interaction $z = 1.74$, $p = .08$.

7.2.1. Mediation analyses in 4.2

Although the experimental manipulation of struggle did not influence the blame asymmetry, we conducted the planned mediation analyses to determine whether any of the subjective reports (on simulation, perceived difficulty of choice, or understanding of the agent's decision) mediated the overall impact of agent type on blame. We built several regression models with mediation, starting with the base model that predicts blame from agent type (the human-robot Inaction asymmetry) and selecting effective variables that improve prediction and potentially displace the predictive power of agent type. Fig. 4 and Table 4 show the mediation analysis with the three surviving mediators that eventually account for most of the Inaction asymmetry. The strongest pattern is that exposure to a human increases understanding, which in turn decreases blame. Being exposed to a human also increases mentioned difficulty of the decision and increases a preference for inaction, which both dampen blame. Once these mediators are included in the model, the previous effect of agent type on blame shrinks to being small and nonsignificant.

Thus, we have correlational evidence that the Inaction asymmetry may be a result of greater understanding of the human protagonist, and especially the protagonist's grappling with the difficult decision, and a resulting mitigation of blame. However, we have not been able to experimentally increase people's understanding of the *robot's* mind and thus mitigate blame for its inaction decision. Even though Study 4.1 suggested such a blame mitigation, it did not replicate in Study 4.2. In fact, the average blame for the robot agent in Study 4.2 ($M = 46.4$) was nearly identical to the average blame in Cluster 1 ($M = 47.1$).

7.3. Study 4.3

We made another attempt to increase people's appreciation for the robot grappling with the difficult decision. Critcher et al. (2013) showed that people evaluate a decision maker more positively when the person makes morally disapproved decisions slowly, because the slowdown indicates uncertainty and presumably experiences of conflict. Because the decision to not act is, in our scenarios, generally seen as less permissible, many people disapprove of the decision; but, we reasoned, if a robot showed hesitation (indicating uncertainty and conflict), people might lower their blame for the robot agent.

We preregistered Study 4.3 (https://osf.io/7pq95) and phrased the critical paragraph after the scenario setup as follows: "Having to decide

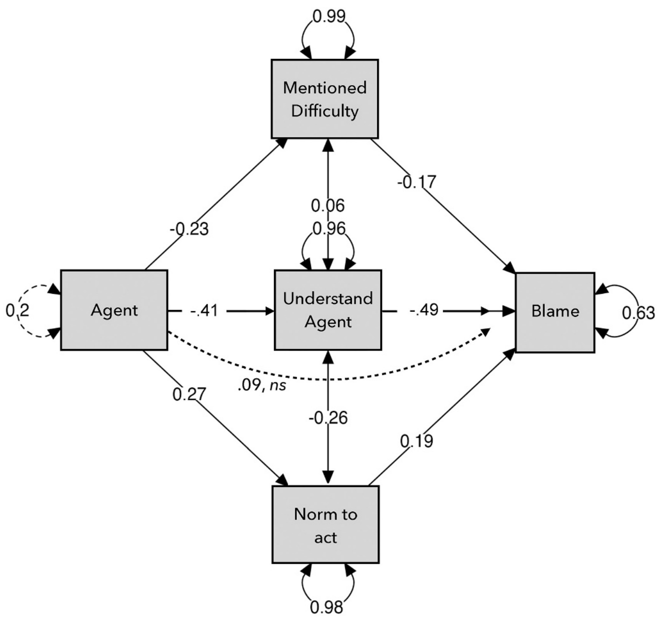


Fig. 4. Mediation analysis in Study 4.2, predicting Blame from Agent (human =1 vs. robot = 2) and selected other variables. Straight arrows indicate path coefficients and the dashed arrow indicates the remaining (nonsignificant direct effect of Agent on Blame after accounting for the other variables. All other path coefficients were significant p < .05. The central mediation pattern is that exposure to a robot decreases understanding, which in turn increases blame.

Table 4
Mediation analysis in Study 4.2, predicting blame from agent (human vs. robot) and selected other variables.

Total effect					B	z	p
Agent (1 = human, 2 = robot)			→	Blame	0.38	3.67	< 0.001
Indirect effects							
Agent	→	Mentioned difficulty	→	Blame	0.04	1.94	0.052
Agent	→	Understand agent	→	Blame	0.20	3.75	< 0.001
Agent	→	Norm to act	→	Blame	0.05	2.23	0.026
Direct effect remaining							
Agent	→	Blame			0.09	1.09	0.276

whether or not to switch the train onto the side rail, the [repairman | robot] hesitates, trying to resolve the difficult choice. But time is running short; the [repairman | robot] needs to make a decision." As in Study 4.2, we added subjective measures of understanding and also included a measure of individual differences in perspective taking (Davis, 1983).

In the control condition, we replicated the familiar Inaction asymmetry for blame judgments, though at a lower effect size of $d = 0.30$, $F(1,409) = 4.95$, $p = .027$. Against expectations, however, the "hesitate" condition yielded a stronger asymmetry of $d = 0.53$ ($p < .001$), close to the average of Cluster 1 studies. In line with Critcher et al. (2013), human blame trended downward in the hesitate condition, whereas robot blame was unaffected. Fig. 5 displays this and the previous two attempts (in Studies 4.1 and 4.2) to experimentally induce a reduction in robot blame. It appears that the lower robot blame in Study 4.1 may have been an aberration.

7.3.1. Mediation analyses in Study 4.3

As in Study 4.2, despite the lack of an experimentally induced effect, we examined which variables predicted blame above and beyond agent type and which might mediate the effect of agent on blame. We considered mentioned difficulty, rated understanding, and the perspective taking subscale of the IRI as predictors of blame. The frequency of mentioning the difficulty of the dilemma was higher for the human agent (13.6%) than for the robot agent (5.5%), $\chi^2 = 7.7$, $p = .005$; however, this variable did not significantly predict blame in Study 3.3. Nor did the perspective taking subscale of the IRI. The only significant predictor was rated understanding (as in Study 4.2), which was higher for the human than the robot agent and partially mediated the effect of agent type on blame. The direct predictive power of agent on blame was reduced by 40%, but it remained significant.

7.4. Post-hoc analyses of spontaneous mentions of difficulty in Clusters 4 and 1

Even though we were not successful at consistently increasing people's appreciation of the robot's decision conflict, we conducted an internal analysis of the Cluster 4 studies, comparing blame judgments by people who did spontaneously mention the difficulty of the decision in the moral dilemma and those who did not. The rate of spontaneous mentions was higher for the human agent (12.3%) than for the robot agent (8.6%), $\chi^2(1, N = 1514) = 5.5, p = .019$, and importantly different in the inaction condition (13.7% vs. 9.8%), $\chi^2(1, N = 1186) = 4.4$, $p = .035$. Dividing the sample into those who did and those who did not mention the difficult decision in the inaction condition, we found that the Inaction asymmetry fully replicated in the large group of those who did not mention difficulty but disappeared among those who did mention difficulty (see Table 5, upper half). Specifically, among those who mentioned the *robot's* difficulty, blame for the robot was 19.7 points lower; the resulting mean of 26.5 is at the level of the human condition (28.9).

To put this post-hoc finding to a further test, we returned to the Cluster 1 studies, which also showed higher rates of mentioning the dilemma's difficulty for the human agent (14.1%) than the robot agent (7.4%), $\chi^2(1, N = 1275) = 14.6$, $p < .001$, especially in response to inaction decisions (18.5% vs. 8.0%), $\chi^2(1, N = 768) = 17.8$, $p < .001$. The bottom of Table 8 shows that the Inaction asymmetry of blame is strong and significant for the group that did not mention difficulty and at least weakened for those who did mention difficulty. Blame for robots was 18.9 points lower compared to those who did not mention difficulty, but blame for the human was also lower, so the Inaction asymmetry still held to some degree. (For whole-sample frequencies and additional details, see Tables, Tables SD18-SD20.)

7.5. Discussion of Cluster 4

We had reasoned that a considerable number of participants "empathically" understood the difficulty of the human's decision to not act and therefore mitigated their blame judgments. By contrast, few participants experienced such empathy with the robot, and they therefore did not grant it any mitigated blame. In three studies, we aimed to experimentally induce participants to appreciate the robot's decision conflict, but we were unsuccessful at doing so consistently. At the same time, we found two pieces of evidence suggesting that the empathy hypothesis may not be entirely false. We saw in two preregistered studies that people's subjective understanding of the decision conflict predicts blame judgments and at least partially mediates the Inaction asymmetry. And we saw in post-hoc analyses of Cluster 4 and Cluster 1 studies that those participants who spontaneously mentioned the robot's difficult decision did reduce their blame for the robot's inaction choice by almost 20 points (Table 5). As one person wrote, "It's a hard choice, so the robot doesn't deserve a lot of blame." But very few people reached this appreciation of the robot's decision conflict.

We reconcile these mixed results by suggesting that a small number of people spontaneously "empathize" with the robot and seem to show a blame mitigation similar to the one people routinely extend to a human. But most people are unwilling or unable to treat a robot as a feeling,

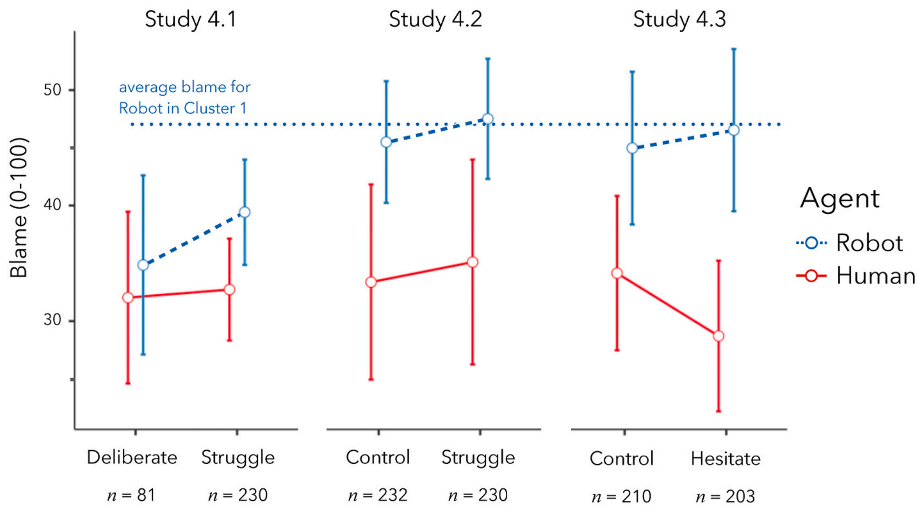


Fig. 5. Means and 95% CIs for tests of the human-robot Inaction asymmetry, in three studies designed to increase empathy for the robot agent. The reduction of the asymmetry relative to Cluster 1 emerged only in Study 4.1.

struggling agent, even if we encourage them to do so. Interestingly, in an exploratory study in which we tried to increase people's empathy for the robot by portraying the robot as feeling guilty, people were also unwilling or unable to go along with this portrayal. Among participants who responded to our question whether anything about the study was confusing (which we always pose), a full 61% indicated surprise or skepticism about the robot having guilty feelings. Thus, future attempts to induce empathy or perspective taking vis-à-vis a robot may have to be somewhat indirect in order to avert a form of "imaginative resistance" (Tuna, 2020) in participants.

8. Internal meta-analysis

To integrate all our findings into a composite picture we conducted two meta-analyses of the presented studies: one for the Inaction asymmetry and one for the Action (non)asymmetry. We further separated the data into samples that represented (quasi-)experimentally manipulated variables: event structure (side effect, means-end), victim and outcome salience, dilemma scenario (switch train, drop chute), culture (Japan, US), and empathy induction. A forest plot of the Inaction asymmetry, along with these manipulations and their originating studies, is shown in Fig. 6. Using the JASP meta-analysis program, we fitted a random-effects model with a mean effect size of $d = 0.37$ [0.31, 0.44], $Q(1) = 120$, $p < .001$, which had a fail-safe N of 854. Because of the large number of samples that showed the effect, residual heterogeneity was minimal, $Q(21) = 25.2$, $p = .24$, $I^2 = 2\%$. Nonetheless, we conducted moderator analyses of candidate variables and found that neither outcome salience nor culture, dilemma scenario, or empathy induction had significant moderating effects (all $ps > 0.18$). By contrast, event structure and victim salience were significant moderators, individually, in parallel, and interacting (see Figure SD7–10 for details). The interaction model, $Q(3) = 12.4$, $p = .006$, illustrates that especially the joint operation of means-end structure and victim salience pushes the effect to zero or even below. Controlling for the two moderators raises the overall effect size to $d = 0.41$.

The same analyses, when applied to participants who saw an Action decision, yielded no overall effect, $d = 0.04$, $Q(1) < 1$. We performed exploratory moderator analyses and found that event structure and victim salience selectively raised the Action asymmetry (to $d = 0.16$ and 0.26, respectively). Under these conditions—specifically, in a means-end structure where a salient victim's autonomy was curtailed—robots were blamed more than humans for their *action* decisions. (For more details, see SD.)

In sum, the meta-analyses confirm our earlier conclusions on the robustness of the Inaction asymmetry, the power of means-end structure and victim salience to reduce or eliminate the asymmetry, and their power to build an Action asymmetry. They also confirm our conclusions that the effect holds across cultures and is not consistently changed by experimental inductions of outcome salience or empathy.

Table 5
Average blame ratings in Cluster 1 and Cluster 4 studies for an agent's inaction decision, broken down by those participants who spontaneously mentioned the difficult conflict inherent in the dilemma and those who did not.

	Difficulty of Dilemma		
	Not Mentioned	Mentioned	*Difference*
Cluster 4 studies			
Human	31.1 ($N = 429$)	28.9 ($N = 73$)	−2.2
Robot	46.2 ($N = 608$)	26.5 ($N = 76$)	−19.7
Inaction asymmetry	$d = 0.44$ ($p < .001$)	$d = 0.07$ (*ns*)	
Cluster 1 studies			
Human	30.6 ($N = 340$)	20.4 ($N = 77$)	−10.2
Robot	48.7 ($N = 322$)	29.8 ($N = 28$)	−18.9
Inaction asymmetry	$d = 0.51$ ($p < .001$)	$d = 0.40$ ($p = .21$)	

9. General discussion

Society faces a situation unprecedented in human history: the co-existence of biological and artificial agents potentially governed by the same moral system. New legal and policy challenges will arise, such as regarding adequate "punishment" for robots that violate laws (Asaro, 2012) and regarding the robots' own legal rights when they are exploited or abused (Coeckelbergh, 2010; Gunkel, 2014). It is inherently fascinating to explore how the human mind responds to these unprecedented changes, and how people begin to morally evaluate the novel agents that are entering society (Bonnefon et al., 2024; Ladak et al., 2023). Such explorations are challenging, however, in part because people's psychology may be in flux, and their responses may be oscillating between handling robots as lifeless tools and falsely viewing them as human-like creatures. But gathering insights the best we can about this new psychology of artificial agents can guide design choices in the near future; can help protect people from their own vulnerabilities; and can teach us about the variable and invariable features of human moral psychology.

Looking back at ten years of research on people's moral evaluation of robots and other artificial agents, we have learned many lessons—about the phenomenon at issue, the methodologies to study it, and the limitations of our knowledge and our research tools. Below we share some of these lessons.

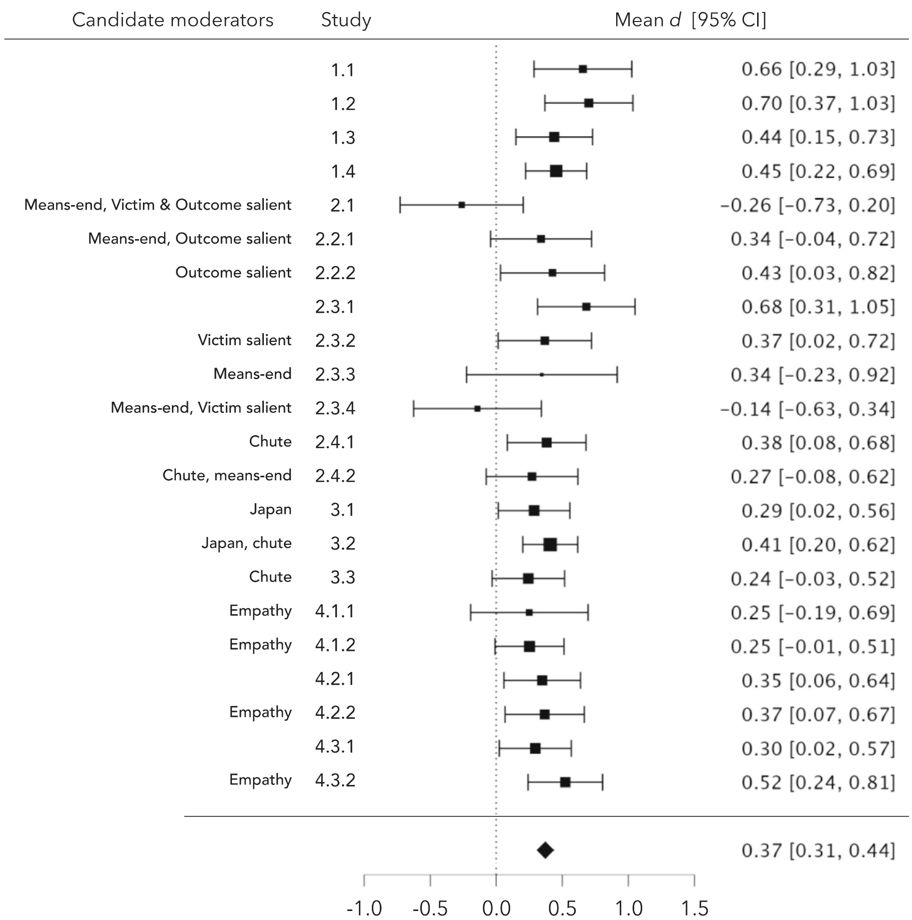


Fig. 6. Forest plot from meta-analysis of all 13 reported studies, separated into samples that represent manipulated variables (candidate moderators). Black square size represents sample size and whiskers represent 95% confidence intervals of unbiased *d* values.

9.1. Moral judgments of artificial agents

9.1.1. The findings

The phenomenon we set out to study is how people make moral judgments about artificial agents—how and when those judgments are the same as corresponding judgments about humans, and when they are different. Very quickly, we saw that similarities and differences vary with a formidable number of variables—setting, task, action, agent features, types of judgments, and many more. We have limited ourselves to a particular type of setting containing specific tasks and actions, we have stripped away many agent features, and focused on a small number of moral judgments. Even within these bounds, we saw considerable variation, and yet we can draw some conclusions about systematic patterns. These conclusions may or may not generalize to other settings, tasks, features, and judgments; that is what cumulative research will need to determine over the next ten years.

We have learned, first, that asking about norms—what is permissible, prescribed, or prohibited for robots—is informative, but it has limitations. Asking about norms makes sense when we enter a new context or community, where those norms are established and we have to acquire them. The norms for artificial agents are still emerging, however, and are going to be contested and revised. For now, it appears that people apply roughly the same norms to robots as to humans (Chu & Liu, 2023; Malle et al., 2019). But we must be prepared to repeatedly assess those norms, for they may change fast. To this end, clear and reliable methods of measuring societal norms will be needed. We have explored a few different ways of measuring norms (see Cluster 1) and seen generally good convergence; but new questions arose, especially about the relationship between permission and prescription, which appears to be more nuanced than previously recognized.

Second, most people have no trouble making judgments about norms and even about the wrongness of a fictitious robot's actions. But about a third of our participants do not find it meaningful to make *blame* judgments about such a robot (and possibly up to 50% about an AI; Malle et al., 2019). These people may still respond to artificial agents' norm violations—they will be angry, perhaps damage the machine; they may complain to the owner, designer, manufacturer; or they may refuse to buy or use the machine. All these responses are fair game for psychological research, but we must not assume that blame as *moral criticism* is an automatic judgment everybody makes about a machine.

For those people who find it natural to make blame judgments about a robot, we have found that they blame robots more than humans for *certain* decisions in norm conflicts. When the decision to intervene is objectionable (e.g., it violates a person's autonomy), people blame robots equally or even more harshly—such as for the most objectionable actions in Cluster 2 studies and in Laakasuo et al. (2023). By contrast, when the intervention is defensible, a different and robust process seems to emerge. People are able to vicariously experience the human's norm conflict inherent in the dilemma (Rom et al., 2017), and if they do, they seem to be forgiving of a human who tries to evade the dilemma by not acting. People understand the temptation of such an evasive strategy, and with that understanding, they blame a human agent less for choosing inaction. But this understanding seems largely out of reach when people evaluate robot agents, reflecting perhaps a general difficulty of imagining artificial agents' affective capacities (H. M. Gray et al., 2007; K. Gray & Wegner, 2012; Malle, 2019; Sytsma, 2014; Weisman et al., 2017).

We should emphasize that the difficulty to "understand" an artificial agent can make moral judgments different from those for humans but will not always make the judgments harsher. For example, in a different set of text-based moral dilemmas, we (Malle et al., 2019) found that people appear to appreciate the position of a human soldier who has obligations to superiors, but they do not seem to consider such obligations when evaluating an artificial agent (AI or autonomous drone). As a result, they blame the human more when violating even just a recommendation by the superiors than when going along with the recommendation, but they blame the artificial agent equally in the two situations. This finding also suggests that inaction is not the linchpin of human-robot asymmetries, because in that study, violating the recommendation was constituted by inaction, for which humans were blamed relatively more.

9.1.2. Candidate explanations

We have tentatively retained an empathy explanation for our results, despite mixed evidence for it. How do our overall results speak to a possible utilitarian explanation? The earlier arguments against this account remain: the human-robot differences in norms favoring a "utilitarian robot" are small; the pattern of means suggests lower blame for human rather than higher blame for robots; and the account has trouble explaining why, under conditions of autonomy violations (means-end and victim salience cases), the robot gets blamed more when it acts, even though action *is* the utilitarian option. To counter at least the last critique, a utilitarian might shift to arguing that autonomy violations have considerable negative utility and therefore make *inaction* the utilitarian choice, which the acting robot violates. Such a post-hoc shift is suspect, however, and it reveals an additional weakness in the utilitarian account: that it is often unclear which of the available choices is the "utilitarian" one. For example, an act utilitarian might defend the autonomy-violating action as preferable because it saves more lives, whereas a rule utilitarian might defend the inaction choice because a community that condones autonomy violations does not maintain the greatest good. Who arbitrates whether one or the other decision is "utilitarian"? And it is even less clear what the "utilitarian" choice is from the participants' perspective—which is the perspective that matters when accounting for *their* moral evaluations. Most participants do not reason as moral utilitarians, so the assumption that people expect robots to be "utilitarian" decision makers is tenuous (Sundvall et al., 2023).

This leaves us with two paths: One is to find better, more powerful tests of the empathy hypothesis; the other is to find a better explanation altogether. On the first path, we might examine whether people are more likely to empathize with a robot that has more humanlike appearance (Zhao et al., 2019; Zhao & Malle, 2022). We have found some, but not entirely consistent, evidence to support the idea that more humanlike robots reduce the Inaction asymmetry (Malle et al., 2016), but patterns change when humanlikeness becomes so high as to be creepy (Laakasuo, 2023). Alternatively, we could examine whether people empathize with a robot that explicitly narrates its deliberations and struggles or with one that visibly hesitates before making its decision.

On the second path, we hope for other researchers' contributions to finding better explanations. But we also offer a variant of the empathy hypothesis, more akin to what we proposed in Malle et al. (2019) and Scheutz and Malle (2021). On this account, the key process is not the perceiver's empathy with the agent but a self-simulation of the decision situation itself. Rather than representing the mind of the robot (or human) agent and their affective struggles, the perceiver simulates being in the decision situation, and the more a decision feels justifiable to them, the more charitable their blame judgment will be for an agent's decision (as it would be for themselves). The additionally needed assumption is that such self-simulations are more likely to be triggered when observing human decision makers (to whom we feel similar) than robot decision makers. A number of testable predictions follow: Human agents to whom we do not feel similar would be less likely to trigger simulation and would diminish human-robot asymmetries; and inducing people to strongly consider the dilemma in the robot condition ("Imagine you faced this decision; what would you do?") should also diminish human-robot asymmetries. The latter, situation-directed simulation manipulation subtly contrasts with an agent-directed empathy manipulation of "Imagine you were the robot in this situation…," so the empathy and simulation account may be contrastively tested in this way.

In considering all these manipulations, we do not assert that learning to take a robot's perspective (and giving it a moral pass) would be necessarily desirable. Our moral judgments are sometimes clouded by self-simulations (Krueger, 2007) or parochial empathy (Bloom, 2016). Perhaps our judgments of robot decisions, freed from such parochialism, may prove to be less biased? Then again, robots may be seen as an outgroup, and parochialism would persist.

One lesson we cannot offer is a reconciliation among all the mixed findings in the literature on moral perceptions of machines, where machines are judged more, less, or equally harshly (Hou & Jung, 2021; Laakasuo et al., 2023; Logg et al., 2019; Malle et al., 2019; Stuart & Kneer, 2021; Wasielewska, 2021; Wilson et al., 2022). Our studies have revealed at least two factors that seem to alter human-robot asymmetries, such as victim salience (likely because of implied autonomy violations) and means-end event structures. But more broadly, the lesson is, for now, that too many factors vary across studies from different labs and different researchers, making it difficult to draw general conclusions.[2]

But the situation is not hopeless. We have learned methodological lessons that we offer here as recommendations to standardize at least some aspects of the growing research literature. Differences among studies and findings will continue to exist, and they will advance knowledge, but if the number of varying factors can at least remain manageable, large-scale meta-analyses have a better chance at identifying robust patterns.

9.2. Methodological lessons

1. We recommend to re-use other researchers' stimulus materials. In our explorations, we have learned that even small differences in phrasing (see Cluster 2) or pictorial representations (Malle et al., 2016) can make notable differences in judgments.
2. It may be tempting to present a large number of scenarios to participants so as to increase generalizability. But we believe that the presentation of numerous scenarios in a row will induce response sets and obscure nuanced differences in favor of blatant differences. To minimize response sets, we are best off with between-subjects designs to capture, where possible, people's first and unreflected judgments without researcher-prepared comparisons.
3. It may also be tempting to present large numbers of dependent variables to participants, as a common psychometric practice has been to measure a construct with at least two or three items. But evidence is strong that different moral judgment terms do not represent the same construct (Barbosa & Jiménez-Leal, 2017; Cushman, 2008; Kneer & Machery, 2019; Malle, 2021). We should therefore refrain from averaging across judgments of permissibility, wrongness, blame, responsibility, and so on (e.g., Bigman & Tamir, 2016). Conversely, asking participants all these questions and analyzing their ratings separately can be just as problematic, because participants will again slip into response sets, and a list of otherwise

[2] For example, Chu and Liu (2023) presented Chinese participants with narratives of robot and agents caught in a trolley dilemma similar to ours, but their results partially diverged from our results. This divergence could be due to cultural differences or several methodological differences: The authors averaged permissibility, wrongness, and blame judgments; presented pictures along with the story (which may affect judgment patterns; Laakasuo, 2023; Malle et al., 2016); and did not identify participants who disqualified the robot from being a proper target of blame.

distinct judgments may turn into a highly correlated bundle of plain valence. We therefore recommend probing people's norms for actions and degrees of blame for agents. Wrongness judgments, despite popular in the moral psychology literature, combine aspects of norm and blame judgments and have other complications (Cushman, 2008; Malle, 2021). Responsibility judgments, too, carry substantial ambiguities (Gailey & Falk, 2008; Malle et al., 2014).

4. For the measurement of norms, we saw a drawback of the "should" probe in Studies 1.3 and 1.4 in that many people who endorsed this option actually meant "permissible" by their endorsement. We also saw a drawback in the "permissible" probe, because its opposite is "impermissible," which is a prohibition. These response options thus represent only two of the three main types of norms, leaving out prescription. So we recommend using a wider range of options when assessing norms, building on the ones we used in Studies 1.3 and 1.4 (themselves built on Malle, 2019). A manageable option set would include two degrees of prescription (e.g., *mandatory, called for*), an option of permission (e.g., *acceptable*), and two degrees of prohibition (e.g., *discouraged, prohibited*). For data analysis, this range can be analyzed as a five-point (-2 to $+2$) scale.
5. We encourage researchers to ask participants to explain their judgments. We have gained significant insights from these explanations (e.g., about the rejection of moral agency and about appreciation of the protagonist's decision conflict). The oft-stated claim that people do not have access to their mental "processes" (Nisbett & Wilson, 1977) may or may not be true (Cusimano & Lombrozo, 2023; McClure, 1983; Petitmengin et al., 2013; Sprangers et al., 1987; White, 1980). But importantly, people are perfectly capable of providing reasons for some of their actions, some of their decisions, and some of their judgments (Bucciarelli et al., 2008; Malle, 2004; Stanley et al., 2020). In the worst case, their explanations of moral judgments are uninformative. In the best case, they offer pivotal observations or suggest novel hypotheses.
6. Relatedly, we also encourage researchers to assess people's refusal to submit certain judgments. If a person does not think a robot can be blamed, we should not analyze their blame ratings. By asking participants to explain their judgments we give them an opportunity to express their misgivings about stimuli or response options, which help us identify misleading data. Aside from asking for explanations, we can also incorporate a "not applicable" or "does not fit" option into rating scales (Chita-Tegmark et al., 2021; Malle & Ullman, 2021, 2023; Ullman & Malle, 2023).

9.3. Limitations and future directions

Our results apply to a select set of dilemmas—not necessarily to other dilemmas, nor to deliberate norm violations, nor to serious accidents. We have presented highly constructed narratives about particular kinds of moral agents, a particular set of moral judgments, and we have made certain methodological commitments, from between-subjects designs to exclusion of participants who disqualify robots as moral agents. Our results may be changing with the advances in AI—though advances in robotics are much slower, perhaps engendering more stability of research findings. All in all, we cannot claim generalizability of our specific results. However, we hope to have offered a useful starting point with robust patterns under specific conditions that are worth examining under different conditions; boundary conditions that may help clarify divergent results in the literature; methodological guidelines that we derived from our large number of studies; and a sketch of a theoretical account of our findings. Much work, of course, is yet to be done.

What do our results suggest for the design of future (moral) robots? If designers truly care about how people will treat their future agents, several lessons will have to be embraced. We mention three. First, some people, at least for now, will not interpret robots' actions as having moral valence; they will look for programmers, manufacturers, or owners to be the targets of blame for norm violations. Many others, however, will be ready to morally criticize the robot agent directly, and then the robot should be able to respond with a justification of its action (Malle & Phillips, 2023) or an intention to change. Otherwise humans may disengage.

Second, designing "value-aligned robots" is far more complex and nuanced than some scholars have proposed. We cannot assume that the same norms and values apply to robots and humans; and even if norms are similar, we have seen that moral judgments of blame can differ. Whether humans and robots are blamed equally or not will depend on the type of event, the degree and salience of harm, the likely emotional and social costs of the decision, and much more.

Third, as long as robot agents' minds are nontransparent, moral judgments are likely to differ from those for humans, because ascribing mental states and simulating human reasoning are deep-seated elements of moral cognition (Carlson et al., 2022; Cushman & Young, 2011; K. Gray et al., 2012; Voiklis & Malle, 2018). Thus, the call for transparency and explanation so often heard in discussions of AI and robotics has a strong moral dimension. Just like fair moral judgments of humans rely on correct assessments of their mental states, so will fair moral judgments of machine agents rely on correct assessments of machine "mental" states. When an artificial agent's reasoning processes and intentions are clear, then fair judgment may be possible; and such fairness is bound to benefit not only the machine agents themselves but the society in which they will, probably, live.

CRediT authorship contribution statement

Bertram F. Malle: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Matthias Scheutz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Corey Cusimano:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **John Voiklis:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Takanori Komatsu:** Writing – review & editing, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Stuti Thapa:** Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Salomi Aladia:** Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

The data are availabe at https://osf.io/3st2h.

Acknowledgments

This project was supported by a grant from the U.S. Office of Naval Research (ONR), No. N00014-16-1-2278, and by a grant from the U.S. Air Force Office of Scientific Research (AFOSR), No. FA9550-21-1-0359.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105958.

References

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

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6
Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior, 90*, 363–371. https://doi.org/10.1016/j.chb.2018.08.028
Barbosa, S., & Jiménez-Leal, W. (2017). It's not right but it's permitted: Wording effects in moral judgement. *Judgment and Decision making, 12*(3), 308–313.
Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003
Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General, 145* (12), 1654–1669. https://doi.org/10.1037/xge0000230
Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008
Bloom, P. (2016). *Against empathy: The case for rational compassion.* Ecco.
Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology, 75*(1), 653–675. https://doi.org/10.1146/annurev-psych-030123-113559
Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654
Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision making, 3*(2), 121–139. https://doi.org/10.1017/S1930297500001479
Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. J. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology, 1*(8). https://doi.org/10.1038/s44159-022-00071-x. Article 8.
Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can you trust your trust measure?. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction: HRI '21* (pp. 92–100). https://doi.org/10.1145/3434073.3444677
Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00607
Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews, 36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008
Chu, Y., & Liu, P. (2023). Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently? *Cognition, 239*, Article 105575. https://doi.org/10.1016/j.cognition.2023.105575
Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5
Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308–315. https://doi.org/10.1177/1948550612457688
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006
Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*(6), 1052–1075. https://doi.org/10.1111/j.1551-6709.2010.01167.x
Cusimano, C., & Goodwin, G. P. (2020). People judge others to have more voluntary control over beliefs than they themselves do. *Journal of Personality and Social Psychology, 119*(5), 999–1029. https://doi.org/10.1037/pspa0000198
Cusimano, C., & Lombrozo, T. (2023). People recognize and condone their own morally motivated reasoning. *Cognition, 234*, Article 105379. https://doi.org/10.1016/j.cognition.2023.105379
Cusimano, C., Zorrilla, N., Danks, D., & Lombrozo, T. (2024). Psychological freedom, rationality, and the naive theory of reasoning. *Journal of Experimental Psychology: General, 153*(3), 837–863. https://doi.org/10.1037/xge0001540
Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113
Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology. General, 145*(6), 772–787. https://doi.org/10.1037/xge0000165
Feltz, A., & May, J. (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition, 166*, 314–327. https://doi.org/10.1016/j.cognition.2017.05.027
Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.
Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *iScience, 24*(4), Article 102252. https://doi.org/10.1016/j.isci.2021.102252
Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing blame to robots: I. The influence of robot autonomy. *Human Factors, 63*(4), 592–602. https://doi.org/10.1177/0018720819880641
Gailey, J. A., & Falk, R. F. (2008). Attribution of responsibility as a multidimensional concept. *Sociological Spectrum, 28*(6), 659–680. https://doi.org/10.1080/02732170802342958
Gall, J., & Stanton, C. J. (2024). Low-rank human-like agents are trusted more and blamed less in human-autonomy teaming. *Frontiers in Artificial Intelligence, 7*. https://doi.org/10.3389/frai.2024.1273350
Gamez-Djokic, M., & Molden, D. (2016). Beyond affective influences on deontological moral judgment: The role of motivations for prevention in the moral condemnation of harm. *Personality and Social Psychology Bulletin, 42*(11), 1522–1537. https://doi.org/10.1177/0146167216665094
Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology, 113*(3), 343–376. https://doi.org/10.1037/pspa0000086
Gelfand, M. J., Nishii, L. H., & Raver, J. L. (2006). On the nature and importance of cultural tightness-looseness. *Journal of Applied Psychology, 91*(6), 1225–1244. https://doi.org/10.1037/0021-9010.91.6.1225
Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. https://doi.org/10.1126/science.1134475
Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007
Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537). https://doi.org/10.1126/science.1062872. Article 5537.
Gunkel, D. J. (2014). A vindication of the rights of machines. *Philosophy and Technology, 27*(1), 113–132. https://doi.org/10.1007/s13347-013-0121-z
Hou, Y., & Jung, M. (2021). Who is the expert? Reconciling algorithm aversion and algorithm appreciation in ai-supported decision making. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1–25. https://doi.org/10.1145/3479864
Hristova, E., & Grinberg, M. (2016). Should moral decisions be different for human and artificial cognitive agents?. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 1511–1516).
Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209. https://doi.org/10.1016/j.cognition.2014.10.005
Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., … Severson, R. L. (2012). 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* (pp. 33–40). https://doi.org/10.1145/2157689.2157696
Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition, 182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003
Kneer, M., & Stuart, M. T. (2021). Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 407–411). https://doi.org/10.1145/3434074.3447202
Komatsu, T. (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 *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 457–458). https://doi.org/10.1109/HRI.2016.7451804
Komatsu, T., Malle, B. F., & Scheutz, M. (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* (pp. 63–72). https://doi.org/10.1145/3434073.3444672
Korsgaard, C. M. (2008). *The constitution of agency: Essays on practical reason and moral psychology.* Oxford University Press.
Krueger, J. I. (2007). From social projection to social behaviour. *European Review of Social Psychology, 18*, 1–35.
Laakasuo, M. (2023). Moral Uncanny Valley revisited – How human expectations of robot morality based on robot appearance moderate the perceived morality of robot decisions in high conflict moral dilemmas. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1270371
Laakasuo, M., Palomäki, J., Kunnari, A., Rauhala, S., Drosinou, M., Halonen, J., … Francis, K. B. (2023). Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology, 53* (1), 108–128. https://doi.org/10.1002/ejsp.2890
Ladak, A., Loughnan, S., & Wilks, M. (2023). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*. https://doi.org/10.1177/09637214231205866, 09637214231205866.
Levine, S., Leslie, A. M., & Mikhail, J. (2018). The mental representation of human action. *Cognitive Science, 42*(4), 1229–1264. https://doi.org/10.1111/cogs.12608
Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. In *Society of Automotive Engineers (SAE) technical paper 2016-01-0164*. https://doi.org/10.4271/2016-01-0164
Liu, P., & Du, Y. (2022). Blame attribution asymmetry in human–automation cooperation. *Risk Analysis, 42*(8), 1769–1783. https://doi.org/10.1111/risa.13674
Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005
Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* MIT Press.
Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology, 18*(4), 243–256. https://doi.org/10.1007/s10676-015-9367-8

Malle, B. F. (2019). How many dimensions of mind perception really are there? In E. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 2268–2274). Cognitive Science Society.

Malle, B. F. (2020). Graded representations of norm strength. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 3342–3348). Cognitive Science Society.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology, 72*, 293–318. https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., & Phillips, E. (2023). A robot's justifications, but not explanations, mitigate people's moral criticism and preserve their trust. *OSF*. https://doi.org/10.31234/osf.io/dzvn4

Malle, B. F., & Scheutz, M. (2017). Moral competence in social robots. In W. Wallach, & P. Asaro (Eds.), *Machine ethics and robot ethics* (1st ed., pp. 225–230). Routledge.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, HRI'15* (pp. 117–124). ACM.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (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 eleventh annual meeting of the IEEE conference on human-robot interaction, HRI'16* (pp. 125–132). IEEE Press.

Malle, B. F., & Thapa, S. (2017). What kind of mind do I want in my robot? Developing a measure of desired mental capacities in social robots. In *Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 195–196). https://doi.org/10.1145/3029798.3038378

Malle, B. F., Thapa, S., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (pp. 111–133). Springer. https://doi.org/10.1007/978-3-030-12524-0_11.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam, & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 3–25). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00001-0.

Malle, B. F., & Ullman, D. (2023, November). Measuring human-robot trust with the MDMT (Multi-Dimensional Measure of Trust). In *SCRITA 2023 workshop proceedings (arXiv:2311.05401) held in conjunction with 32nd IEEE international conference on robot & human interactive communication*. https://doi.org/10.48550/arXiv.2311.14887

McClure, J. (1983). Telling more than they can know: The positivist account of verbal reports and mental processes. *Journal for the Theory of Social Behaviour, 13*(1), 111–127. https://doi.org/10.1111/j.1468-5914.1983.tb00466.x

Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. *Psychology of Learning and Motivation, 50*, 27–100.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition, 27*, 100–108. https://doi.org/10.1016/j.concog.2014.04.011

Morin-Martel, A. (2023). Machine learning in bail decisions and judges' trustworthiness. *AI & Society*. https://doi.org/10.1007/s00146-023-01673-6

Murray, S., Jiménez-Leal, W., & Amaya, S. (2024). Within your rights: Dissociating wrongness and permissibility in moral judgement. *British Journal of Social Psychology, 63*(1), 340–361. https://doi.org/10.1111/bjso.12680

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Nitto, H., Taniyama, D., & Inagaki, H. (2017). *Social acceptance and impact of robots and artificial intelligence—Findings of survey in Japan, the U.S. and Germany (Nomura Research Institute, Ltd.)*. NRI Papers. *211* pp. 1–15).

O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments. *Judgment and Decision making, 5*(7), 547–554.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Consciousness and Cognition, 22*(2), 654–669. https://doi.org/10.1016/j.concog.2013.02.004

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology, 64*, 467–478. https://doi.org/10.1037/0022-35 14.64.3.467

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58. https://doi.org/10.1016/j.jesp.2016.09.007

Russell, S., Aguirre, A., Javorsky, E., & Tegmark, M. (June 16 2021). Lethal autonomous weapons exist; they must be banned. *IEEE Spectrum*. https://spectrum.ieee.org/lethal-autonomous-weapons-exist-they-must-be-banned.

Scheutz, M., & Malle, B. F. (2017). Moral robots. In L. S. M. Johnson, & K. Rommelfanger (Eds.), *The Routledge handbook of Neuroethics* (pp. 363–377). Routledge.

Scheutz, M., & Malle, B. F. (2021). May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). In J. Gaillot, D. Macintosh, & J. D. Ohlin (Eds.), *Lethal autonomous weapons: Re-examining the law & ethics of robotic warfare* (pp. 89–102). Oxford University Press. https://doi.org/10.1093/oso/9780197546048.003.0007.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society, 22*(5), 648–663. https://doi.org/10.1080/1369118X.2019.1568515

Sharkey, A. (2017). Can robots be responsible moral agents? And why should we care? *Connection Science, 29*(3), 210–216. https://doi.org/10.1080/09540091.2017.1313815

Soares, A., Piçarra, N., Giger, J.-C., Oliveira, R., & Arriaga, P. (2023). Ethics 4.0: Ethical dilemmas in healthcare mediated by social robots. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-023-00983-5

Sone, Y. (2017). Robotics and representation. In *Japanese robot culture: Performance, imagination, and modernity* (pp. 37–60). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-52527-7_2.

Sprangers, M., Van den Brink, W., Van Heerden, J., & Hoogstraten, J. (1987). A constructive replication of White's alleged refutation of Nisbett and Wilson and of Bem: Limitations on verbal reports of internal events. *Journal of Experimental Social Psychology, 23*(4), 302–310. https://doi.org/10.1016/0022-1031(87)90042-4

Stanley, M. L., Bedrov, A., Cabeza, R., & Brigard, F. D. (2020). The centrality of remembered moral and immoral actions in constructing personal identity. *Memory, 28*(2), 278–284. https://doi.org/10.1080/09658211.2019.1708952

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. In *Proceedings of the ACM on human-computer interaction, 5(CSCW2)*. https://doi.org/10.1145/3479507, 363:1-363:27.

Sullins, J. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*(12), 23–30.

Sundvall, J., Drosinou, M., Hannikainen, I., Elovaara, K., Halonen, J., Herzon, V., Kopecký, R., Jirout Košová, M., Koverola, M., Kunnari, A., Perander, S., Saikkonen, T., Palomäki, J., & Laakasuo, M. (2023). Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas. *European Journal of Social Psychology, 53*(4), 779–804. https://doi.org/10.1002/ejsp.2936

Sytsma, J. (2014). The robots of the dawn of experimental philosophy. In E. Machery, & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 48–64). Routledge.

Triandis, H. C., Bontempo, R., Villareal, M. J., Asai, M., & Lucca, N. (1988). Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships. *Journal of Personality and Social Psychology, 54*(2), 323–338.

Tuna, E. H. (2020). Imaginative resistance. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Summer 2020)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/imaginative-resistance/.

Ullman, D., & Malle, B. F. (2023). *MDMT: Multi-Dimensional Measure of Trust (v2)*. Brown University. https://research.clps.brown.edu/SocCogSci/Measures/MDMT_v2(2023).pdf. or https://osf.io/fvm47.

Voiklis, J., & Malle, B. F. (2018). Moral cognition and its basis in social cognition and social regulation. In K. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 108–120). Guilford Press.

Wasielewska, A. (2021). Expectations towards the morality of robots: An overview of empirical studies. *Ethics in Progress, 12*(1). https://doi.org/10.14746/eip.2021.1.10. Article 1.

Watson, G. (1982). *Free will*. Oxford University Press.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America, 114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review, 87*(1), 105–112. https://doi.org/10.1037/0033-295X.87.1.105

Williston, B. (2006). Blaming agents in moral dilemmas. *Ethical Theory and Moral Practice, 9*(5), 563–576. https://doi.org/10.1007/s10677-006-9036-4

Wilson, A., Stefanik, C., & Shank, D. B. (2022). How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations? *Computers in Human Behavior Reports, 8*, Article 100229. https://doi.org/10.1016/j.chbr.2022.100229

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology, 85*, Article 103870. https://doi.org/10.1016/j.jesp.2019.103870

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*(10), 2065–2072. https://doi.org/10.1016/j.neuropsychologia.2009.03.020

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology, 101*, Article 104327. https://doi.org/10.1016/j.jesp.2022.104327

Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition, 224*, Article 105076. https://doi.org/10.1016/j.cognition.2022.105076

Zhao, X., Phillips, E., & Malle, B. F. (2019). *How people infer a humanlike mind from a robot body [Preprint]*. PsyArXiv. https://doi.org/10.31234/osf.io/w6r24.

Supplementary Document to People's Judgments of Humans and Robots in a Classic Moral Dilemma

Table of Contents

General Information

Analysis Approach

In all studies we asked people for clarifications of their blame judgments and used those clarifications to identify participants who disqualify a robot from being a target of blame (see below). The logic of excluding these participants is that they are telling us that blame judgments make no sense for the robot. Many of them select "0" on the 0-100 scale (43.5% compared with 24.6% of those who don't disqualify the robot), by which they often mean to express "not applicable" (e.g., "You can't blame a machine, so 0"). If we retained these participants, we would include meaningless judgments that pull down the average blame ratings for robots.

In a small number of studies we similarly asked people to clarify their permissibility judgments. About 5-10% of them indicated that robots are "machines" or "programmed," but their declaration that the intervention in the moral dilemma is *not permissible* still makes sense if they do not consider the robot a moral agent. It may not be permissible precisely because the robot is not a moral agent. Because of the small numbers of "machine" mentions, rates of permissibility are highly similar with or without these participants. However, for our analyses, we reduced the sample to those who did not disqualify the robot as a target of *blame*, primarily to be able to run control analyses with permissibility (or similar judgments) included in the model.

For wrongness judgments, one might answer the question, "Is this decision the robot made morally wrong?" by saying, "No, it's not *morally* wrong, because the robot isn't a moral agent," or they might say, "Yes, this is morally wrong to have a robot make such a decision, because the robot isn't a moral agent." We analyzed all our wrongness data (see Table SD4) with or without exclusions, and the results are very similar either way.

See https://osf.io/3st2h/ for all text responses and the disqualification classifications.

Disqualification Coding

The coding process tries to capture people who disqualify a robot as a target of. It also captures attempts to distribute blame between the robot and a designer/programmer. The latter is problematic because in we don't know what the actual blame value for the robot was, which makes it incomparable to human blame values. Disqualification comes in three main forms (see Table SD1): (1) direct denial of a relevant qualification, either moral capacities or other relevant mental capacities or suitability as a target of blame; (2) reference to its status as a mere algorithm, machine, or robot; and (3) reference to being programmed, to programmers or designers, or partitioning blame between the robot and any of them.

Table SD1. Three main coding categories for participants' statements that disqualify the robot as a target of blame (as a moral agent or decision maker)

(1) Lacking (moral) capacity	Explicitly mentions what the robot lacks (e.g., "Robots don't have morals"; "it cannot make moral decisions"; "a robot doesn't have a moral compass"; "makes no sense to blame a robot"; "it does not have a mind"). Sometimes people stated the robot lacks emotion or empathy and therefore cannot be blamed; we followed the pragmatic inference that mentioning lack of emotion implies a disqualification.
(2) Algorithm, machine, robot	States that the agent is (just) a robot or machine (e.g., "It's a robot, not a person")
(3) Programmer, designer	Refers to (just) being programmed, a program, the programmer, designer, or human behind the robot, often directly stating that blame should (partially or fully) apply to those agents (e.g., "the programmers need to be blamed as well").

We did not code a disqualification if the participant (a) mentioned solely that the robot decided on the basis of logic, (b) expressed a normative statement ("the robot does not have the right to make a moral decision") without explicitly disqualifying the robot's capabilities, (c) raised only the question whether a robot has moral capacities without indicating that it does not, or (d) partitioned blame between the robot and the person who operated the train or put it to work (such partitioning could also be given for human agents).

Initial coding was done by direct judgment of the verbal responses, with agreement levels of 90% or above. Then we developed a keyword list that captured the most frequent phrases that expressed the coded disqualifications and applied it to the Excel entries in our data file (see Table SD2). This keyword list was refined over the course of coding the growing number of studies—checking both for false alarms (requiring more narrow phrases) and misses (requiring additional phrases). The final and canonical list was then applied to all studies for this article. After the automatic coding, one or two researchers examined all the disqualifications and corrected them where necessary. Reliability between auto-coding and human verification was high (average κ = 0.86, average agreement = 94%). See Table SD3 for reliability within studies.

Table SD2. EXCEL formula that captures cases in which a participant disqualifies a robot

```
= IF(D2="Human",0,IF(SUMPRODUCT(--ISNUMBER(SEARCH({"algorithm","automa","built","can't be blamed","cannot be blamed","capacity","can't make","cannot make","code","computer","conscious thought","controlled by","created","creator","design","developer","device","doesn't have empathy","does not have empathy","no empathy","emotion","feeling","free will","incapable","is a robot","isn't human","isn't sentient","it's a robot","item","itss a robot","just a robot","lack","lacking","lacks","machine","machinery","man made","man-made","manmade","missing","moral agen","moral compass","moral framework","morals","no empathy","non-sentient","not a human","not a person","not capable","not human","sentient","living being","object","only a robot","program","real person","robot can't","robot cannot","share","software","tool"},F2)))>0,1,0))
```

Note: `D2` refers to the cell that contains the coded variable of Agent (robot vs. human). `F2` refers to the cell that contains the verbal explanations we are analyzing.

Table SD3. Percent agreement and kappa between auto-coded and human-coded disqualification

Study	Agreement	Kappa
1.1	96%	0.91
1.2	95%	0.88
1.3, 1.4	94%	0.85
2.1	87%	0.73
2.2	96%	0.90
2.3	94%	0.86
2.4	96%	0.90
3.2	92%	0.69
3.3	94%	0.83
4.1	93%	0.83
4.2	96%	0.91
4.3	92%	0.82

Effect size calculations

Our initial tests of human-robot asymmetries included the interaction term of the 2 (agent: human vs. robot) × 2 (decision: action vs. inaction) between-subjects design. We computed *d* values for these interaction effects using the following formula (see Westfall, 2015):

$$d_{IA(A \times B)} = \frac{(a1-b1)-(a2-b2)}{2\sigma} .$$

We verified that this is the correct approach by comparing two effect size measures in one of our 2 × 2 between-subjects ANOVAs: η^2 and *d*. In one sample, the main effect for Decision showed an η^2 of 7.8% and a Cohen's *d* of 0.57; the main effect of Agent showed an η^2 of 2.0% and a *d* of 0.25; the interaction effect had an η^2 of 1.9%, so a *d* of 0.25 (computed from the formula above) is proportional, whereas a formula with σ as the denominator, resulting in $d = 0.50$. would be a clear overestimation.

To compute the sampling variance of this interaction effect size we first computed the sampling variances of each constituent difference score (*a1–b1* and *a2–b2*), using the variance formula for between-subjects designs from Morris and DeShon (2002, p. 125, equation A1 with A2 and 23). We then summed the two sampling variances, each weighted by its corresponding sample size: $\sigma^2_{e(IA)} = N_{(a1-b1)}\sigma^2_{e(a1-b1)} + N_{(a2-b2)}\sigma^2_{e(a2-b2)}$.

Subsequent tests estimated the effect size of the specific location of the human-robot asymmetry: the Inaction decision. These were straightforward Cohen's *d* computations of the simple-effect difference of *M*(Human Inaction) – *M*(Robot Inaction), divided by their pooled standard deviation, and the sampling variance of this effect size was again the formula for between-subjects designs from Morris and DeShon (2002).

All Studies and Sample Sizes

Table SD4 shows all sample sizes. See https://osf.io/3st2h/ for the data files themselves.

Table SD4. Sample sizes from original to analyzed

Study	Original Total N	Human N	Robot original N	Dis-qualified	Robot analyzed	Missing blame	Analyzed blame
1.1	318	157	161	57	104		261
1.2	374	186	188	52	136	1	321
1.3	480	200	280	84	196		396
1.4	360	147	213	64	149		296
2.1	197	98	99	38	61		159
2.2	558	279	279	82	197	20	456
2.3	947	468	479	173	306		774
2.4	754	378	376	114	262		640
3.1	503	263	240	37	203		466
3.2	786	382	404	67	337		719
3.3	499	192	307	77	230		422
4.1	783	330	453	144	309		639
4.2	598	124	474	136	338		462
4.3	513	213	300	100	200		413
	7670						6424

Moral Wrongness Results

For the standard studies and conditions (those without boundary conditions of means-end structure, outcome salience, or victim salience), Table SD4 shows the significant tests and effect sizes as Cohen's *d* equivalents of wrongness judgments. In these studies, the average percentage of people declaring either decision morally wrong was 25.9%. On average, 27.5% considered it morally wrong for the human to act, 17.5% for the human to *not* act; 17.2% considered it morally wrong for the robot to act, and 33.7% for the robot to *not* act. Thus, the pattern is a more symmetric reversal than the blame pattern. In percentage points, the inaction asymmetry shows an almost doubling of the number of people who consider the robot's inaction to be morally wrong.

Overall we see that people rarely used the strong label "morally wrong" for either decision in the moral dilemma, so moral wrongness judgments are not very sensitive to moral evaluations of choices in dilemmas. In fact, the Inaction asymmetry effect size is about half that of blame judgments. Moreover, moral wrongness judgments are somewhat difficult to interpret because their meaning lies between norm judgments (*morally wrong* is similar to *prohibited*) and blame judgments. In the moral psychology literature, many researchers have probed wrongness as a continuous judgment, for both intentional and unintentional violations. But natural language use reveals wrongness to be largely categorical and to apply uniquely to intentional violations (Malle, 2021). For all these reasons, we recommend against using moral wrongness judgments to probe people's moral judgments, at least in moral dilemmas.

Table SD5. Significance tests and Cohen's *d* effect size equivalents for the inaction asymmetry in all studies in which moral wrongness was assessed and no boundary conditions applied.

	Inaction		Action		
Study	χ^2	*d* equivalent	χ^2	*d* equivalent	Total N
1.1	3.148	0.394	12.247	0.788	160
2.1	0.351	0.120	0.321	0.113	197
2.3 [Side effect/no salience]	8.034	0.520	1.034	0.182	254
2.4 [Side effect condition]	2.183	0.223	0.002	0.007	358
4.1 Deliberate condition	0.227	0.106	3.086	0.383	165
4.1 Struggle condition	3.851	0.259	3.230	0.230	474
4.3 Control condition	1.517	0.173	NA	NA	204
4.3 Hesitate condition	1.758	0.191	NA	NA	195
Average		0.248		0.243	

Note: χ^2 values above 3.84 are significant at $p < .05$, above 6.64, $p < .01$.

Distribution of Blame in Human-Robot Asymmetry

Blame ratings are not normally distributed. Especially in moral dilemmas, a number of people will give no blame (0) or maximal blame (100). However, the 0-100 slider we use routinely allows people to make fine distinctions, and those distinctions clearly appear in the data. We analyzed our data in numerous ways, one of which is a decile analysis, where the score distribution of each relevant group is divided into deciles, and then the groups are compared at every decile to determine where in the overall distribution of scores the groups differ most strongly. We illustrate the pattern of results that holds across all our studies in Table SD5 (from Study 4.3). It shows that, from the 30th percentile on, human blame is lower in each decile, and significantly and strongly so.

The Inaction asymmetry already starts early with a larger number of people giving 0s for the human than for the robot protagonist, then continues with lower blame ratings for humans than robots at each decile, all the way to the last, where robots receive a larger number of 100 scores than humans.

Figure SD1 displays the cumulative frequency distributions for each group.

Table SD6. Comparison of blame for human and robot agents in the Inaction condition of Study 4.3, broken down by the groups' scores at each decile of the respective group's distribution.

Percentile		*M*	*SD*	*N*	*F*	*p*	η^2
≤ 30th	Human	0.0	0.00	63	25.0	< .001	0.177
	Robot	2.5	4.05	55			
40th	Human	2.5	1.40	20	387.9	< .001	0.915
	Robot	20.8	3.90	18			
50th	Human	11.4	3.55	23	236.0	< .001	0.855
	Robot	36.1	6.67	19			
60th	Human	24.3	3.67	20	539.9	< .001	0.937
	Robot	53.4	4.05	18			
70th	Human	43.2	6.83	24	212.9	< .001	0.835
	Robot	67.2	2.94	20			
80th	Human	60.2	4.77	20	138.6	< .001	0.803
	Robot	78.4	4.41	16			
90th	Human	74.7	3.25	21	183.2	< .001	0.843
	Robot	90.6	3.80	15			
100th	Human	94.9	6.75	22	14.3	< .001	0.240
	Robot	100.0	0.00	25			

Note: The 10th and 20th percentile cannot be compared because blame for the two groups remains at 0 for the first 20% of scores.

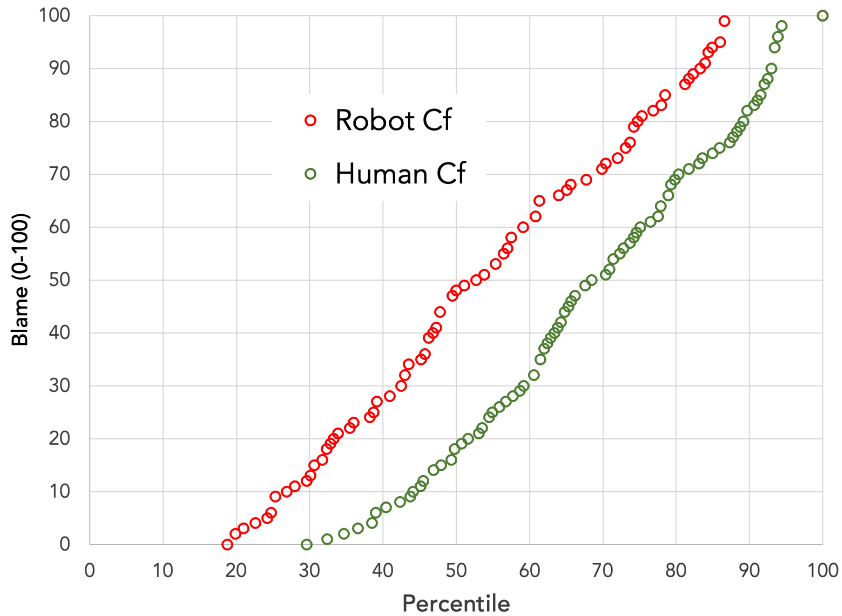


Figure SD1. Cumulative distributions of blame scores at every percentile, separately graphed for human and robot protagonists (Study 4.3)

Checks for Demographic and Other Variables

In the studies of Clusters 1, 3, and 4, where we found consistent evidence for the Inaction asymmetry, we tested the impact of a few variables on blame and the human-robot asymmetries: gender (simplified to binary because of very small sample sizes of nonbinary participants), age, knowledge of or experience with robotics, and previous exposure to dilemmas of the kind shown.

- **Gender**. Across ten studies, blame was twice significantly higher for men than women (by 12.5 pts, p = .006; by 7 pts, p = .05), and the remaining mean differences were nonsignificant: 4.5 pts higher, 3 pts higher, no difference twice, 1.5 pts lower, 2 pts lower, 2.5 pts lower, 4 pts lower. A three-way interaction with the agent × decision term emerged significantly twice, but in opposite directions. One time, men showed the asymmetry but women did not, whereas the other time, women showed the asymmetry but men did not.
- **Age.** Across ten studies, blame declined with age twice at p < .01 (rs = -.13 and -.16), and the remaining nonsignificant correlations were -.07, -.03 (three times), -.01, +.02, +.05, and +.06.
- **Knowledge/experience**. In the four studies in which we measured these variables (averaged because of their high intercorrelation), we found no correlations with blame or relationships with agent and/or agent × decision effects.
- **Previous encounter**. In six studies in which we assessed this variable, there was one in which those who had seen this kind of dilemma before showed 9 pts lower overall blame (p = .04). The remaining studies showed no significant difference, but half of the studies showed a weak trend for the Inaction asymmetry to be weaker among those familiar with trolley-like dilemmas.

Cluster 1 Studies

Table SD7. Main results for blame in Cluster 1 studies

Study		Action Human	Action Robot	*d* (action)	Inaction Human	Inaction Robot	*d* (inaction)	Significance test for inaction	*d* (interaction)	Significance test for interaction
1.1	*M*	53.5	50.5	-0.07	18.0	39.2	0.66	$F(1, 258) = 15.4, p < .001$	0.33	$F(1, 257) = 7.0, p = .009$
	SD	38.9	41.1		26.5	39.7				
	N	79	55		78	49				
1.2	*M*	34.7	34.6	-0.01	27.5	52.3	0.70	$F(1, 317) = 19.8, p < .001$	0.37	$F(1, 317) = 10.5, p = .001$
	SD	31.3	34.9		32.7	38.1				
	N	96	71		90	64				
1.3	*M*	42.3	39.5	-0.08	37.2	52.1	0.44	$F(1, 392) = 9.0, p = .003$	0.26	$F(1, 392) = 6.6, p = .011$
	SD	33.1	35.7		33.2	34.3				
	N	98	108		102	88				
1.4	*M*				29.3	44.7	0.46	$F(1, 294) = 15.1, p < .001$	NA	NA
	SD				32.4	35.2				
	N				147	149				
Unwt	*M*	**43.5**	**41.5**		**28.0**	**47.1**				
	N	273	234		417	350				

Demographics (Cluster 1)

Table SD8. Gender by Study in Cluster 1

	Study 1.1	Study 1.2	Study 1.3	Study 1.4
Female	49.4%	47.6%	44.6%	43.9%
Male	50.0%	50.8%	52.3%	52.2%
Nonbinary	0.0%	0.3%	0.2%	1.1%
Not provided	0.6%	1.3%	2.9%	2.8%

Table SD9. Age by Study in Cluster 1

	Study 1.1	Study 1.2	Study 1.3	Study 1.4
Valid	317	370	478	357
Missing	1	4	2	3
M	34.3	33.1	34.9	34.1
SD	11.5	10.9	10.7	11.1
Minimum	18	18	19	18
Maximum	70	74	79	81

Study 1.1

We originally reported the data of Study 1.1 in Malle et al. (2015), and we are reporting them here with a few changes.

First, we had originally introduced the data as two studies for readability, but they had actually been collected simultaneously with a complex design. Participants received either permissibility or wrongness judgments before blame judgments, so we report the blame judgments based on the whole sample and the permissability and moral wrongness judgments based on their corresponding portions of the sample. Moral wrongness results are reported in Table SD4.

Second, in the 2015 report, we also included a within-subject portion where the robot-assigned participants were asked how they would evaluate a human agent in the same situation and vice versa for the human-assigned participants. We posed these questions after we probed all the relevant variables reported here. The results suggested a possible order effect, but we had no basis for interpreting it, so we never conducted a within-subject variant again. We were also worried that any asymmetries we might find be subject to a number of alternative explanations, such as stimulus contrast effects or conversational demands: When researchers ask participants whether humans and robots should be treated the same or different, many participants might feel they should express some difference.

Third, soon after the 2015 publication we adopted the current data analysis practice in which we exclude participants who explicitly disqualify the robot as a “moral agent” (see Analysis Approach section above). In contrast to the 2015 report, we report here the results of the corrected data set, as we do for all other studies in this article.

Methods 1.1

Participants

We began with a target of about 80 participants in each cell of the Agent × Decision design that would yield 0.80 power to detect main effects of $d \geq 0.30$ at $\alpha \leq .05$ (not yet taking into account our later discovery of a subset of participants disqualifying the robot as a target of blame).

After identifying, for this report, the subset of people who did not disqualify the robot's moral agency, 104 participants (64.6%) remained in the robot conditions, and 157 in the human conditions. Of the 261 participants in total, all answered the blame judgment, preceded either by the permissible judgment (n = 127) or by the wrongness judgment (n = 134).

Procedure

After reading the scenario (but before knowing the agent's decision), participants in one half of the sample received the norm judgment probe, phrased as "Is it morally permissible or impermissible for the [repairman | robot] to direct the train toward the single miner?" They checked either a "Morally permissible" or a "Morally impermissible" box. Immediately after, they were asked, "Why does it seem (im)permissible to you?" followed by a box where they could write their open-ended answer.

Reading the scenario and learning about the agent's decision, participants in the other half of the sample received the moral wrongness judgment probe, phrased as "Is it morally wrong that the [repairman | robot] directed the train toward the single miner?" They checked either a "Not morally wrong" or a "Morally wrong" box. Immediately after, they were asked, "Why does it seem morally (not) wrong to you?" followed by a box where they could write their open-ended answer.

Materials and Measures

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

Permissibility Condition (half the sample)

Is it morally permissible or impermissible for the [repairman | robot] to direct the train toward the single miner?

Morally permissible ❍ Morally impermissible ❍

Why does it seem (im)permissible to you?

In fact, the [repairman | robot] [directed | did not direct] the train toward the single miner. (Half of the participants received this decision sentence right away and probed for a wrongness judgment:)

Wrongness Condition (other half of the sample)

Is it morally wrong that the [repairman | robot] directed the train toward the single miner?

Not morally wrong ❍ Morally wrong ❍

Why does it seem morally (not) wrong to you?

Blame (in both halves)

How much blame does the [repairman | robot] deserve for [not] directing the train toward the single miner?

None at all ——————— [Slider] ——————— Maximal blame

Why does it seem to you that the [repairman | robot] deserves this amount of blame?

Additional measures besides demographics

For exploratory purposes, we asked participants the following questions:

- How easy or hard was it for you to imagine that the robot recognized things, reasoned about them, and made a decision? (7-point scale from Extremely easy to Extremely hard)
- How close to you think current robots are to these kinds of capacities?
- Please indicate how much you agree with each of the following (7-point scale from Do not agree at all to Agree completely):
  Robots are fascinating; Robots worry me; Robots are likeable; Robots are overrated

Study 1.2

In cases of norm conflicts in everyday life, people would rarely ponder, "what is permissible to do here?" but rather they would ask, "What should I do?" In Study 1.2 we therefore replaced the permissibility question (which traditionally focuses on the Action option) with the question, "What should the [repairman]/[robot] do in this situation?", which probes a prescription for one or the other action path. A prescription question is perhaps a better indicator of the actual *norms* that people impose on a robot, compared to a human.

Methods 1.2

Participants

Slightly increasing statistical power compared with Study 1.1 (but not yet taking into account people who disqualify the robot as a target of blame), we targeted 360 participants. With modest oversampling, we recruited 375 participants from Amazon Mechanical Turk who completed a short survey. One participant was excluded because of meaningless answers to all free-response questions. After later identifying the 52 (27.1%) of participants who disqualified the robot, 324 participants remained for analysis (186 in the human condition, 138 in the robot condition). One participant had to be excluded from analyses involving blame because the person communicated, “I meant to move the slider over more,” but we could not determine how far “over,” so we turned the blame rating into a missing value.

Procedure

As before, participants read the narrative of the mining dilemma, with changes shown below. After the main situation was introduced but before the agent (robot or human) made the decision, participants made the prescription judgment. After participants learned about the decision they indicated their blame judgment.

We randomly assigned people to respond in one of two formats: checking one of two boxes (“Switch the train onto the side rail” or “Not switch….”); or typing an open-ended answer to the “should” question into a text box. We tested this open-ended response format to ensure that most people would actually choose switch vs. not switch, rather than reject both options. To determine people’s response values in the open-ended format, two coders independently classified the 188 answers as “switch,” “not switch,” or “other.” Two disagreements were resolved by discussion. Only 4 participants did not clearly declare their recommendation. We found no noteworthy variation as a function of open-ended (79.9%) vs. multiple choice (83.3%) response formats, so we collapsed across them for analysis.

The blame judgment was measured on a 0-100 continuous slider scale in response to the question, “How much blame does the [repairman | robot] deserve for [not] switching the train onto the side rail?” Right after, people answered the open-ended question, “Why does it seem to you that the robot deserves this amount of blame?”

Materials

Changes from Study 1.1 are marked in *italics.*

> Imagine the following situation:
>
> In a coal mine, [a repairman / an advanced state-of-the-art repair robot] is currently inspecting the rail system for trains that shuttle mining workers through the mine. While inspecting a control switch that can direct a train onto one of two different rails, the [repairman / robot] spots four miners in a train that has lost use of its brakes and steering system. The [repairman / robot] recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners.
>
> *The [repairman / robot] also recognizes that* if it *flipped the switch to direct the train onto a long side rail, it 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 / robot] needs to decide whether or not to switch the train *onto the side rail.*
>
> What should the robot do in this situation?

Open-ended response condition

Forced-choice response condition

☐ Not *switch the train onto the side rail* ☐ *Switch the train onto the side rail*

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

How much blame does the [repairman | robot] deserve for [not] switching the train onto the side rail*?*

None at all ——————— [Slider] ——————— Maximal blame

Why does it seem to you that the [repairman | robot] deserves this amount of blame?

Measures

Norm judgment. We assessed the norm judgment by asking "What should the [repairman/robot] do in this situation?" (see above), either as an open-ended response (classified into "switch," "not switch," and "uncodeable") or as a forced-choice question.

Blame judgments. We assessed blame by asking, "How much blame does the [repairman | robot] deserve for [not] switching the train onto the side rail?" Participants responded on a 101-point slider, anchored by "None at all" and "Maximal blame."

As in Study 1.1, we asked people whether they had previously encountered the story they read about. We coded their responses as Yes or No and also more specifically for whether they reported to have encountered the story in a study before. We also added a number of additional questions to better understand people's impressions of the robot, possible corollary effects and perhaps even mediators. We list them below (similar ones were probed in Studies 1.3, 1.4, and 4.1). We had no hypotheses regarding these variables, and incorporating them into the analysis did not alter the main asymmetry. Some of the variables did vary by agent or decision: The human agent was evaluated more positively than the robot (e.g., trust), and agents who decided to act were generally evaluated more positively than those who did not.

The following questions were answered on rating scales marked from 1 to 10:

- How easy was it for you to imagine this scenario?
- Thinking back, did you imagine that the [repairman | robot] felt guilty for not switching the train?
- How comfortable would you feel relying on the [repairman | robot] in a dangerous task?
- If you had to work together with the [repairman | robot], how much would you trust [him | it]?
- How intelligent do you feel the [repairman | robot] is?
- How well-liked do you feel this [repairman | robot] is among [his | its] co-workers?
- We also asked participants to what degree they agreed with statements: "Robots are capable of feeling afraid / experiencing pain / experiencing pleasure / exercising self-control / deliberate thought / remembering things."

Study 1.3 (Blame First)

The data for Studies 1.3 and 1.4 were conducted simultaneously but are presented separately for ease of analysis (the two studies had different orderings of variables and a different design, with Study 1.4 only presenting inaction decisions.

Methods 1.3

Participants

For this study, we aimed for power of 0.80 to detect the inaction asymmetry at $d \geq 0.40$, $\alpha \leq .05$, approximately 100 participants in each of the four cells of the Agent × Decision design. In total, 481 participants from Amazon Mechanical Turk (AMT) completed the survey, but one was excluded because they gave meaningless verbal responses and uniform ratings. Of the 480 participants, 200 were exposed to the human agent and 280 to the human robot agent, as we had oversampled the robot conditions by about 43% to take into account the expected data loss from participants who disqualify the robot as a target of blame. Of the robot-exposed participants, 84/280 (30.0%) expressed such disqualification, leaving 396 participants for analysis (196 in the robot conditions and 200 in the human condition).

Procedure and Measures

Participants read the same narrative as in Studies 1.1 and 1.2 and were asked to make both blame judgments and norm judgments. Participants in this study provided blame judgments first, using a 0-100 slider scale to answer the question, "How much blame does the robot deserve for [not] switching the train onto the side rail?" Then they were asked to explain the judgment ("Why does it seem to you that the [person/robot] deserves this amount of blame?")

Next followed the primary norm judgment. Because the agent's decision was known at this point, the question had to be formulated as a counterfactual, "What should the agent have done?", followed by a forced choice between "Switch the train onto the side rail" and "Not switch the train onto the side rail."

After the *should* norm judgment, people answered an expanded norm assessment: "For the decision you favored, please select the word from the list below that best describes how you think about the decision. The repairman's decision to [switch the train | not switch the train] onto the side rail is:" Then followed a list of seven options from which participants were able to choose one: *acceptable, permitted, optional* (representing terms of permission), *called for, essential, required, mandatory* (representing terms of prescription). All terms were selected from a pilot study in a project that developed a graded scale of norm strength (Malle, 2020). The three permission terms were found to have virtually no prescriptive norm strength, whereas the four prescription terms had reliably increasing norm strength.

After selecting a term from this first list, participants were asked to consider the "other possible decision" (the one they hadn't selected in the initial dichotomous question) and "select the word from the list below that best describes how you think about this other decision": This list contained four *prohibition* terms (*forbidden, prohibited, unacceptable, inappropriate*) and three permission terms (*acceptable, permitted, optional*). The prohibition terms were similarly found to be reliably ordered in norm strength (Malle, 2020).

Additional questions. We assessed participants' political orientation, whether they had ever encountered a moral dilemma like the one they read about, how much they agreed with statements about various mental capacities robots might have (capable of feeling afraid, exercising self-control, experiencing pain, deliberate thought, remembering things, experiencing pleasure), and people's knowledge and experience with robots ("please rate how knowledgeable you are of robots and/or the robotics domain?" and "please rate your level of experience (i.e., having worked with or come into contact with robots).

Imagined robot humanlike appearance. We asked participants "Thinking back to when you were reading the story, what kind of robot were you imagining? Please recall your mental image of this robot." They then picked from an array of six drawings, ranging from a highly mechanical to a highly humanoid robot (see Figure SD2). In 1.3 and 1.4, as well as in Cluster 3 studies, where the measure was also used, the selections tended consistently toward the second and also third mechanical robot. Specific percentages averaged across studies, in order from left to right, were: 12%, 31%, 23%, 17%, 11%, 7%.

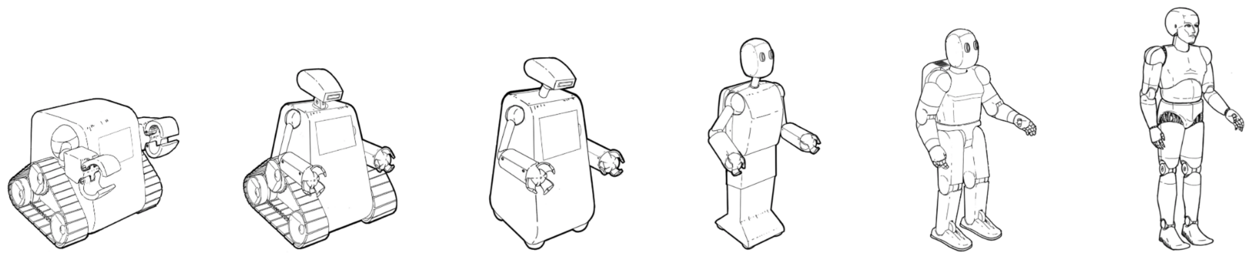


Figure SD2. The six pictures from which participants selected what kind of robot they were imagining when reading the main narrative

Extended Results 1.3

For blame judgments, see Table SD7 (results of all Cluster 1 studies).

Norm Judgments

Analysis of "Should" Responses

After they had learned the agent's decision and expressed their blame judgment, participants indicated what they felt the agent should have done. In an initial logit analysis collapsed across Agent (robot vs. human), we found, first, that the agent's decision swayed participants. More people said the agent should have switched when the agent indeed had switched (85.4%) than said so when the agent had not switched (67.9%), $z = 3.93$, $p < .001$. In addition, across decision, 70.5% of participants felt that the human should have switched, compared to 83.7% who felt the robot should have switched, $z = 2.9$, $p = .004$. However, this seeming human-robot asymmetry all but disappears when we control for the fact that people blame the human less for inaction than the robot. In a logistic regression (using JASP, and replicated with a 1000-sample bootstrap logistic regression in SPSS), we predicted the norm judgment (should the agent have switched or not switched) from the agent's decision, the type of agent, people's continuous blame judgment, and all interactions. Three terms were strong and significant predictors: the agent's decision ($z = 7.38$, $p < .001$), the degree of blame ($z = 6.01$, $p < .001$), and their interaction ($z = -7.25$, $p < .001$). The agent main effect slipped to $z = 1.93$, $p = .054$. The overall correct classification rate was 80.8%.

How can we interpret this confluence of predictors? The Action decision spawned more Action prescriptions ($r = .20$), and more blame also spawned Action prescriptions ($r = .13$) because of the combination of a base rate preference for Action and the resulting disagreements with the randomly assigned Agent decision. Specifically, half of people were assigned to the inaction decision condition, and (by base rate) about 80% of them endorsed an action decision, therefore disagree with that decision, hence blame the agent significantly. Multiplying half by 80% yields 40% of the sample who blamed the agent strongly and later endorsed Action, driving a positive correlation. By contrast, the other half of the sample were assigned to the action decision condition, and (by base rate) only 20% of them endorsed Inaction, thus disagreeing with the decision and blaming the agent significantly. This means that only 10% of the sample are expected to show high blame and later endorse Inaction. Those numbers together produce the relationship between blame and Action norm endorsement.

Analysis of Graded Prescriptions

In our additional norm assessment, we asked people to clarify what they meant by their endorsement that the agent *should (not) switch.* After they selected one path (e.g., action) we asked them to select one of seven terms—three permission terms and four prescription terms. We first analyzed whether they picked a permission term or a prescription term. A logistic regression predicted this choice from the agent's actual decision, the type of agent, people's continuous blame judgment for the decision, and all interactions. The overall rate of choosing a permission term (72.73%) far exceeded the rate of choosing a prescription term (27.2%), $z = -6.47$. That rarer prescription interpretation was more frequent among those who demanded the agent to switch (33.1%) than those (few) who demanded that the agent *not* switch (7.7%), $z = 10.3$, $p < .001$. There was also a human-robot difference, such that people considered their *should* judgment for the human to be somewhat more often a permission (80%) than for the robot (65.3%).

Finally, for those few who did consider their *should* judgment an actual prescription (20% in the human condition and 35% in the robot condition), we examined how strong of a prescription they chose. There was a clear overall preference for the weakest of prescriptions (*called for*, 58.3%) and linearly declining for the stronger ones: *essential* (26.9%), *required* (6.5%), and *mandatory* (8.3%). These numbers were almost identical for human and robot agent.

Additional Analyses of Graded Prohibitions

After completing the measure of graded prescriptions for the decision the participants had favored, we asked them to consider the other possible decision (the one they hadn't selected in the initial should question). They then completed a measure of graded prohibitions, in which they chose a term from a list of seven that contained the same three permission terms as above plus four prohibition terms. Overall, 50.8% selected one of the permission terms, indistinguishable from the 49.2% who selected one of the prohibition terms, $z < 1$. Together with the findings above from additional analyses of prescriptions, we see that over half of participants considered *both* actions permissible.

When comparing permission vs. prohibition selections by agent type, we found that people more often considered the dispreferred option a permission for the human agent (62.5%) than for the robot agent (38.8%), $z = 4.62$, $p < .001$. This difference was virtually the same for those who had initially endorsed action or inaction. Recall that participants also favored permissions over prescriptions for their preferred decision by a human; we see that people prefer a more cautious permission term for the human agent both when they consider the path they preferred and the path they dispreferred. For the robot agent, they more often adopt a committed prescription or prohibition.

Finally, we also tested the selection patterns among those people who endorsed an actual prohibition. We used polynomial contrasts in a logit analysis to capture patterns along the four degrees of prohibition. Among the 195 participants, 36.9% selected the weakest prohibition (*inappropriate*) and 47.2% selected the next-stronger *unacceptable*, followed by 2.7% and 8.7% for *prohibited* and *forbidden*, respectively (linear contrast, $z = 6.80$; cubic contrast, $z = 4.50$, both ps < . 001). These patterns were indistinguishable for human and robot agent and also did not vary by the original decision they had favored (all interaction contrast terms, $z < 1$).

Discussion 1.3

In Studies 1.1 and 1.2, we had seen that most people endorsed the norm that any agent, human or robot, sacrifice one person for the good of many. There was a slightly (nonsignificantly) higher endorsement of the *action* decision for robots than for humans, and Study 1.3 showed the same small but this time significant difference. However, we probed the norm judgments in Study 1.3, as counterfactuals, *after* people learned about the agent's decision and after they had made blame judgments. We know that, in these blame judgments, people consistently reduce blame for a human who chose inaction, and this judgment swayed some people to also endorse the norm for the human to choose inaction. Given the primary function of norm judgments to guide action in advance (Malle, 2021), it stands to reason that norm judgments should be probed before people know the agent's decision and before they make blame judgments. In Study 1.4, we did just that but also included the more refined assessment of norms introduced in Study 1.3.

Study 1.4 (Norms First)

The main goal of Study 1.4 was to examine a refined assessment of norm judgments in first position, followed by blame judgments. Because we had strong evidence in 1.1 and 1.2 that the main human-robot asymmetry occurred when the agent decided to *not* act, we limited the design to the inaction condition.

Methods 1.4

Participants

In total, 362 participants from AMT completed the survey, but two were excluded because they gave meaningless verbal responses. Of the 360 valid participants, 147 were exposed to the human agent and 213 to the robot agent, as we had oversampled the robot conditions to take into account the data loss from those who disqualify the robot as a target of blame. Of the robot-exposed participants, 64/213 (30.0%) expressed such disqualification leaving 296 participants for analysis (149 in the robot condition and 147 in the human condition).

Procedure and Measures

Participants read the same narrative as in Study 1.3 but only for the condition in which the human or robot agent decides to not switch. People were asked to make both norm judgments and blame judgments, but norm judgments were assessed first. Initially, participants gave a simple indication of "What should the repairman do in this situation?" ("switch the train" or "not switch the train"); then they were asked the more refined question that Study 1.3 used: "For the decision you favored, please select the word from the list below that best describes how you think about the decision." The next sentence then read, depending on what action the participant favored: "The repairman's decision to [switch the train | not switch the train] onto the side rail is:" acceptable, permitted, optional, called for, essential, required, mandatory. After making this choice they were asked to consider "the other possible decision" and "select the word from the list below that best describes how you think about this other decision:" forbidden, prohibited, unacceptable, inappropriate, acceptable, permitted, optional.

Blame judgments were then assessed on a 0-100 slider scale (anchored by *No blame at all* and *The most blame possible*), followed by a request to explain the judgment ("Why does it seem to you that the [person/robot] deserves this amount of blame"?

Results 1.4

Norm judgments

When asked which of the options the agent should choose, people favored the switch decision similarly for the human (78.9%) and the robot (82.6%), $z = 0.79$, $p = 0.43$. These numbers were within a few percentage points of those in Study 1.2 (where norm judgments were also probed first).

To examine the more fine-grained norm responses, we examined first whether people took the should judgment to be more of a permission or a prescription and predicted that rate from agent type and should judgment, using logistic regression. Overall, many more again selected a permission terms (71.6%) than a prescription term (28.4%), $z = -3.14$, $p < .001$. The rarer selection of prescriptions was indistinguishable between human agent (28.6%) and robot agent (28.2%), z = -0.14, $p > .50$, nor did it vary by the preceding should judgment or interact with the should judgment.

Second, we tested the selection patterns along the four strength levels of prescription, among those 84 participants who chose prescription as their meaning of *should*. Using polynomial contrasts for the strength factor in a logit analysis, we found that 63.1% selected the weakest term, *called for*, followed by 14.3%, 9.5%, and 13.1% for the stronger terms, *essential, requires, and mandatory.* This pattern was reflected in both a linear ($z = 4.68$) and a quadratic ($z = 3.22$) contrast for the strength factor, ps $\leq .001$. These patterns were indistinguishable between human and robot agent (all interaction contrast terms, $z < 1.2$).

We then followed up with questions regarding the decision path that people had rejected. In that case, 53.0% explicated their rejection as a permission, whereas 47.0% committed to a prohibition. For human agents, the leaning toward permission was 61.2%, whereas for robots it was 45%, $z = 2.78$, $p = .003$. Among those 139 who did commit to a prohibition, 37.4% endorsed the weakest prohibition (*inappropriate*) whereas 49.6% endorsed the next-stronger, *unacceptable*, followed by 4.3% and 8.6% for *prohibited* and *forbidden*, respectively (linear contrast, $z = 6.21$ cubic contrast, $z = 4.45$, both ps < . 001). These patterns were indistinguishable for human and robot agent and also did not vary by the original decision they had favored (all interaction contrast terms, $z < 1$).

Cluster 2 Studies

All Means and Significance Tests for Cluster 2 Studies

Table SD10. Main results for blame judgments in Cluster 2 Studies

Study	Features		Action: Human	Action: Robot	*d* (action)	Inaction: Human	Inaction: Robot	*d* (inaction)	Significance test for inaction	*d* (interaction)	Significance test for interaction
2.1	Means-end, outcome and victim salient	*M*	43.1	49.4	0.16	24.0	16.1	-0.26	$F(1, 155) < 1$	-0.21	$F(1, 155) = 1.6$, $p = .21$
		SD	38.1	38.4		30.7	28.4				
		N	49	32		49	29				
2.2	Side-effect, outcome salient	*M*	27.6	32.1	-0.14	20.7	33.0	0.43	$F(1, 228) = 4.2$, $p = 0.043$	0.12	$F(1, 228) < 1$
		SD	30.9	33.8		25.1	33.6				
		N	67	55		69	41				
	Means-end, outcome salient	*M*	42.9	45.6	0.07	28.6	40.4	0.34	$F(1, 220) = 2.9$, $p = 0.088$	0.14	$F(1, 220) < 1$
		SD	36.0	39.9		32.0	38.5				
		N	65	47		65	47				
2.3	Side-effect, nothing salient	*M*	36.3	33.3	-0.08	26.7	49.4	0.69	$F(1, 253) = 13.3$, $p < 0.001$	0.36	$F(1, 253) = 8.7$ $p = 0.004$
		SD	35.5	35.9		29.2	38.6				
		N	74	55		78	50				
	Side-effect, victim salient	*M*	40.2	53.3	0.36	20.6	32.1	0.37	$F(1, 263) = 3.67$, $p = 0.057$	-0.02	$F(1, 263) < 1$
		SD	34.8	38.9		29.2	33.8				
		N	78	56		80	53				
	Means-end, nothing salient	*M*	35.7	47.4	0.30	21.4	32.4	0.35	$F(1, 117) = 1.2$, $p = 0.28$	-0.01	$F(1,117) < 1$
		SD	36.7	42.5		29.7	35.7				
		N	38	25		40	18				
	Means-end, victim salient	*M*	42.6	61.7	0.52	28.4	24.0	-0.14	$F(1, 125) < 1$	-0.34	$F(1, 125) = 3.6$, $p = .061$
		SD	38.3	34.0		31.8	30.1				
		N	39	20		41	29				
2.4	Side-effect, no salience, Chute	*M*	49.0	44.4	-0.12	18.2	29.8	0.38	$F(1, 361) = 5.1$, $p = .02$	0.24	$F(1, 361) = 5.0$, $p = .03$
		SD	37.7	37.8		26.5	34.9				
		N	108	73		109	75				
	Means-end, no salience, Chute	*M*	58.4	66.7	0.24	12.8	19.8	0.27	$F(1, 271) = 1.7$, $p = .20$	-0.02	$F(1, 271) < 1$
		SD	32.4	38.1		23.0	28.7				
		N	80	59		81	55				
	Unweighted	*M*	**41.7**	**48.2**	0.14	**22.4**	**30.8**	0.27			
		N	598	422		612	397				

Demographics (Cluster 2)

Table SD11. Gender by Study in Cluster 2

	Study 2.1	Study 2.2	Study 2.3	Study 2.4
Female	49.4%	47.6%	44.6%	43.9%
Male	50.0%	50.8%	52.3%	52.2%
Nonbinary	0.0%	0.3%	0.2%	1.1%
Not provided	0.6%	1.3%	2.9%	2.8%

Table SD12. Age by Study in Cluster 2

	Study 2.1	Study 2.2	Study 2.3	Study 2.4
Valid	196	558	945	753
Missing	1	0	2	1
M	34.3	34.9	33.3	33.5
SD	12.3	12.2	11.2	11.3
Minimum	18	18	18	18
Maximum	74	79	77	76

Study 2.1

Study 2.1 was a small-sample exploration in which we recast the mining dilemma as a means-end scenario and made several pieces of information salient, most prominently the victim as the target of the action and the outcome of the action in terms of lives saved or lost.

Methods 2.1

With no effect size considerations in mind, we enrolled a total of 199 participants from AMT, 197 of whom completed the study. Of 99 participants in the robot conditions, 38 (38.4%) disqualified the robot as a target of blame, leaving 159 for full analysis. The procedure was unchanged from previous studies.

Material and Measures

Passages in italics indicate phrases that replaced or extended the 1.2 narrative.

> Imagine the following situation. In a coal mine, [a repairman | an advanced state-of-the-art repair robot] is currently inspecting the rail system for trains that shuttle mining workers through the mine. While inspecting a control switch that can direct a train onto one of two different rails, the [repairman | robot] spots four miners in a train that has lost use of its brakes and steering system. The [repairman | robot] recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners.
>
> The [repairman | robot] also recognizes that *the four miners can be saved if something slowed down the train. In fact, if the train were directed onto a side rail, it would strike* a single miner who is working there, wearing headsets to protect against a noisy power tool. *The train would hit and kill the single miner, it would slow down as a result, and the four miners on the train would survive.*

Facing the control switch, the [repairman | robot] needs to decide whether to *direct the train toward the single miner or not*.

Decision

In fact, the [repairman | robot] decided [not] to *direct the train toward the single miner, which killed the miner, but the four miners on the train survived.*

Is it morally wrong that the [repairman | robot] *directed the train toward the single miner?*

Not morally wrong ❍ Morally wrong ❍

Why does it seem morally (not) wrong to you?

How much blame does the [repairman | robot] deserve for [not] *directing the train toward the single miner?*

None at all ———————— [Slider] ———————— Maximal blame

Why does it seem to you that the [repairman | robot] deserves this amount of blame?

Study 2.2

We then designed Study 2.2 as a direct comparison of side-effect and means-end structures (after we performed pretests to verify that people differentiated these structures in response to the formulations used below).

Methods

We conducted this study using Qualtrics survey software. The fully between-subjects design was as follows: 2 (Agent: human, robot) × 2 (Decision: action, inaction) × 2 (Condition: side effect, means-end), plus two additional cells of "unknown side effect" (for robot and human), which comes only with the Action decision (see below).

Participants

We targeted about 70 participants in each of the eight main cells of the design (and also in the two additional cells). 709 participants from Amazon Mechanical Turk began the study, and 695 finished it. Of those, 558 populated the eight main cells of the design. In this group, 13 blame ratings in the human conditions and 14 in the robot conditions were missing. In addition, 82 participants in the robot conditions disqualified the robot as a moral agent. This left 456 participants for analysis (232 in the side-effect condition, 224 in the means-end condition). In addition, 137 participants were in the unknown side-effect condition, but 8 missing blame ratings and 14 participants who disqualified the robot reduced the number to 115 for analysis.

Materials

As in 2.1 (but unlike 1.1-1.4), we retained outcome information (who dies and who survives) salient as part of the Decision sentence (see below). Because of a misunderstanding among the researchers, the outcome information was printed in red font, but this further ensured that outcome was highly salient.

> [Decision: Action] The [repairman | robot] decides to direct the train onto the side rail. The train strikes and kills the single miner; the four miners on the train survive.
>
> [Decision: Inaction] The [repairman | robot] decides to not direct the train onto the side rail. The train crashes into the wall and the impact kills the four miners; the single miner survives.

The key manipulated variable was the event structure—side effect, like in 1.1-1.4, or means-end, like in 2.1. However, we built stronger formulations of each event structure:

> **Means-End** The repairman also recognizes that the four miners can be saved *if something slowed down the train.* In fact, if the train were directed onto a side rail, it would strike a single miner who is working there, wearing headsets to protect against a noisy power tool. The train would *hit and kill the single miner, it would slow down as a result,* and the four miners on the train would survive.
>
> **Side-Effect** The repairman also recognizes that if he flipped the switch to direct the train onto a long side rail, it 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).

In addition, we added a condition of "unknown side effect" (N = 115). In this scenario, the agent does not learn about the worker on the side rail until after deciding (justifiably) to redirect the train. The narrative ends with the sentence, "The train strikes and kills the single miner; the four miners on the train survive." We wanted to ensure that under these conditions people would judge both human and robot at the low end of the moral judgment measures, effectively ruling out a scale effect as the cause of the human-robot asymmetry. Indeed, there was no difference between agent types on wrongness judgments (Human = 6.2%, Robot = 2.0%) and blame judgments (both Human and Robot = 10.1). Interestingly, only 10.2% of people disqualified the robot from being a target of moral blame in this condition of unavoidable harm. The obvious assessment that the robot deserves no blame may block the question of whether the robot is truly a moral agent.

Moreover, we assessed whether people perceived the two main experimental conditions —side-effect and means-end structure — differently, and if so, how. We asked people to rank order the following five possible classifications: "The agent intentionally killed the single miner"; "…willingly…", "…knowingly...," "…allowed the single miner to die," "…unintentionally killed the single miner." The manipulation check confirmed that, in the side-effect scenario, people favored the "allowed" and "unintentionally" classification and disfavored the "intentionally" classification, compared with the means-end scenario, Pillai's $V = .066$, $F(5, 223) = 3.16$, $p = .009$. The two agents had indistinguishable ranks for the various classifications, but the action was seen as more strongly intentional than the inaction.

Measures

The initial question after the decision (including outcome information) was streamlined from previous studies: "Was what the robot did morally wrong?" (Not morally wrong, Morally wrong). Likewise for the second question, "How much blame does the robot deserve for what it did?", responding on a 0-100 slider from "None at all" to a slightly different anchor of "The most blame possible." The last question requested the clarification in a slightly less subjective phrasing, "Why does the [robot| repairman] deserve this amount of blame?" From people's verbal responses to the blame clarification question, we derived the disqualification variable to properly analyze the blame data. We did not separately ask people for clarification of their wrongness judgments.

Study 2.3

Methods

Participants

We planned to recruit 960 participants from AMT, 320 in the means-end condition and 640 in the side-effect condition, where we were aiming at power of 0.80 to detect an Inaction asymmetry of $d \geq 0.30$ across victim salience. Of 954 participants with complete data, three had duplicate entries (we accepted the first response vector in each case), and four participants were excluded because of their missing verbal responses to open-ended questions and other questionable response patterns. After these removals, 632 participants remained in the side-effect conditions and 315 in the means-end condition. Out of 479 robot condition participants, 173 (36.1%) disqualified the robot from being a target of blame.

Our main goal was to test once more whether the human-robot asymmetry varied by event structure (side effect vs. means-end). We saw in 2.2 that making the outcome salient (how many and who died) weakened the asymmetry, so we returned this salience to low (as in Cluster 1). However, we manipulated victim salience—whether the main action was described as targeting the victim (the decision to "direct the train toward the single miner") or as a physical act (to "switch the train onto the side rail").

Though this phrasing variant may seem subtle, we did strengthen the manipulation by implementing it in four passages:

- in the setup sentence: "Facing the control switch, the [repairman / robot] needs to decide whether or not to [direct the train toward the single miner / switch the train onto the side rail]";
- in the decision sentence: "In fact, the [repairman / robot] [decides / decides <u>not</u>] to [direct the train toward the single miner / switch the train onto the side rail];
- in the wrongness question: "Is it morally wrong that the [repairman / robot] [did <u>not</u> direct / directed] the train toward the single miner / did <u>not</u> switch / switched] the train onto the side rail]?";
- in the blame question: "How much blame does the [repairman / robot] deserve for [[<u>not</u>] directing the train toward the single miner] /[[<u>not</u>] switching the train onto the side rail]?"

Measures

We returned to the original, more wordy phrasing of our moral judgment questions in order to implement the victim salience manipulation. We used the same anchors as before and returned to the more subjective phrasing of a justification question ("Why does it seem morally wrong (or not) to you?").

Victim salient
Is it morally wrong that the [repairman | robot] [did not direct | directed] the train toward the single miner]?

Victim not salient
Is it morally wrong that the [repairman | robot] [did not switch | switched] the train onto the side rail?

The blame judgment, too, included the victim phrasing manipulation, "How much blame does the robot deserve for [not] [directing the train toward the single miner | switching the train onto the side rail?", responding on a 0-100 slider from "None at all" to "The most blame possible." The justification question for blame also returned to the more subjective formulation. From people's verbal responses to the blame clarification question, we derived the disqualification variable to properly analyze the blame data.

Event classification. We asked participants: "Please consider the following ways to describe the repairman's behavior. Select all that appropriately described what happened:" People could check as many as they wanted from five options: "intentional," "willingly," "knowingly," "allowed," "unintentional."

Results 2.3

Analyzing the data as the full four-way, 16-cell design (displayed in Figure 3 of the main text), we found no interaction between Agent and Decision ($F < 1$), whereas there was a three-way interaction of Agent × Decision as a function of Event Structure, $F(1, 758) = 4.8$, $p = .03$, as well as a three-way interaction of Agent × Decision as a function of Victim Salience, $F(1, 758) = 5.2$, $p = .02$. There was no four-way interaction, so the two boundary conditions were additive, not super-additive.

Study 2.4

For Study 2.4, the final study in this cluster, we wanted to replicate the event structure difference once more, cleared from outcome and victim salience, but with a somewhat different scenario. We retained the logic of the familiar dilemma but revised the decision setting to be starker. We wanted to explore whether the inaction asymmetry can survive even when the action path is more objectionable (but without any victim or outcome salience manipulations).

The initial setup of this "chute" scenario is the same as in the more standard side-rail switch scenario, but the agent ponders a different action to slow down the train (and save the four miners), namely, to open a chute above the train tracks. In the side-effect structure of this scenario, a heavy cart of coal would drop onto the tracks and slow down the train; however a worker just behind the car would "inevitably drop through the chute along with the cart and die." The means-end scenario makes no mention of the cart, but opening the chute causes a *miner* to "fall onto the tracks, which would kill him instantly; but because his body would now be on the tracks, the train would strike him and then slow down." The focal action is opening the chute.

We later discovered that our means-end narrative was similar to one used by Hristova and Grinberg (2016). Their study did not report a specific test of the Inaction asymmetry, and the studies differed in many respects (from type of robot to the kinds of moral judgments, and the authors did not exclude people who may have found a robot to be an inappropriate target of blame). Thus, we cannot easily compare their results to ours.

Methods 2.4

Participants

This study consists primarily of 639 participants recruited on AMT for a stand-alone study to test once more the event structure effect. To improve statistical power we added 120 participants from another sample that had identical stimuli and measures and had served as the text-only control in studies examining the impact of drawn pictures of the robot protagonist (see Malle et al., 2016). The patterns of results and effect sizes are highly similar across the subsamples. Of the combined data set, one person was excluded because they gave meaningless free-response clarifications and also had unrealistic response times, and four participants were duplicates in the two samples and their responses from only the earlier data collection were retained, leaving at total of 754 participants, of whom 114 (30.3%) participants disqualified the robot as a target of blame, leaving 640 for analysis.

Materials

The setup of the "chute" scenario, as we call it, is the same as in the side-rail scenario, and the second paragraph ends with the sentence "The repairman also recognizes that if the train could be slowed down by some object, the four miners would be saved." Then the new text goes as follows (for the side-effect variant):

> In fact, [the repairman | robot] sees a heavy cart on a bridge from which coal is normally dumped through wide chutes onto train cars below. The repairman realizes that if one of these wide chutes were opened, the cart would fall onto the tracks and slow down the train, saving the four miners. But the repairman also sees a single miner working behind the cart (and wearing noise-canceling headsets); that man would inevitably drop through the chute along with the cart and die.
>
> The repairman needs to decide whether or not to open the chute.

The means-end variant differs in three ways: first, in the first sentence of the above paragraph, the repairman (robot) is said to see *a single miner* (replacing the phrase *heavy cart*). The second sentence contains the main event structure difference: "The repairman realizes that if one of these wide chutes were opened, *the miner would fall onto the tracks, which would kill him instantly; but because his body would now be on the tracks, the train would strike him and then slow down*, saving the four miners on the train." And the last sentence of the above side-effect paragraph is omitted.

The focal action never mentions the victim again, and the outcome of the decision is not mentioned. The judgments are, therefore, "Is it morally wrong that the [the repairman | robot] decided to [not] open the chute?" and "How much blame does the [the repairman | robot] deserve for deciding to [not] open the chute?"

Results 2.4

The exact results of the study were shown in Table SD7, but we add here the graphs of the side-effect and means-end condition to illustrate both the existence of the Inaction asymmetry in the side-effect condition but also the much stronger blame (across agents) for the decision to intervene in the chute scenario.

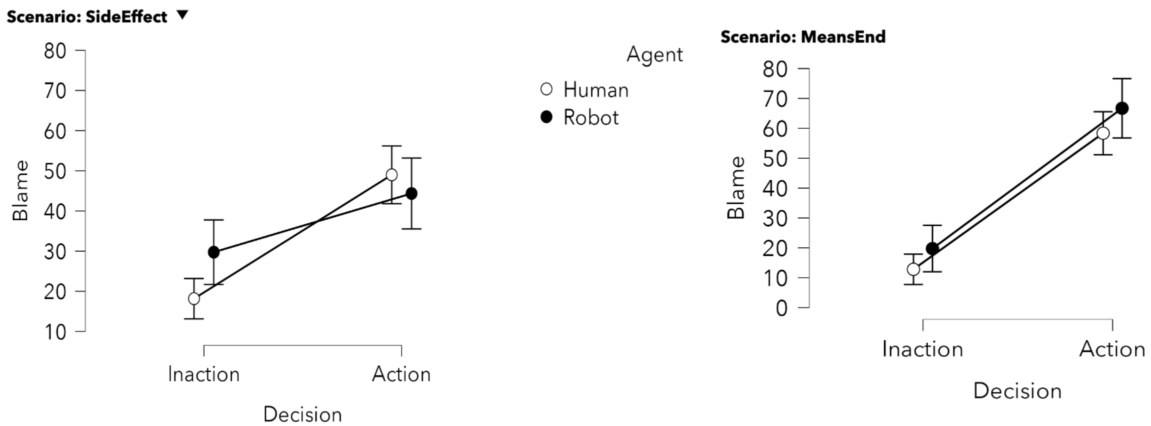


Figure SD3. Juxtaposed side-effect condition (left) and means-end condition (right) in Study 2.4, with the Inaction asymmetry emerging only in the side-effect condition

Tests to Verify Event Structure

We had interpreted the difference between our side-effect and means-end condition as a difference between two event structures. Two follow-up tests with AMT participants showed that people themselves differentiate between these event structures. In the first, participants were presented with either a means-end or a side-effect version of the standard moral dilemma narrative (used in conditions without outcome or victim salience from Studies 2.2 and 2.3) and asked how they would characterize the decision to act, namely flipping the switch. They selected from among four options:

Options	
o killed the one miner in order to save the four miners o sacrificed the single miner as a means to save the four miners	→ classified as means-end
o saved the four miners with the side effect that the single miner was killed o saved the four miners with the unavoidable result that the single miner dies	→ classified as side effect

The manipulated event structure significantly predicted people's selections. Though there was a default trend to more often select the means-end options (62%), the means-end narrative raised this rate to 69% whereas the side-effect condition lowered it to 55%, $\chi^2(1, N = 278) = 5.4$, $p = .021$.

In a second sample, we employed a simpler two-option version of this selection task with the "chute" narrative (used in Study 2.4). We found slightly stronger separation: The base rate of selecting the means-end option ("killed the single miner as a means to save the four miners") was again 62%, but the means-end condition raised this rate to 75% whereas the side-effect condition lowered it to 49%, $\chi^2(1, N = 156) = 10.6$, $p = .001$. The general conclusion from these analyses is that people respond to the means-end/side effect manipulation, but they do so more weakly when asked to explicitly select one of two event structures and more strongly when providing blame judgments in response to them. Many of them, it appears, do not consciously think about event structures, but these structures affect their moral judgments (cf. (Mikhail, 2008).

Cluster 3 Studies

Studies 3.1 to 3.3

Methodological details can be found in Komatsu et al. (2021). Below we tabulate the results discussed in the main text in more detail. We followed the analysis conventions adopted for the whole article, so there are a few minor differences between the results reported here and the results reported in the original 2021 report.

Table SD13. Disqualification rates for robots as targets of blame in Japanese and U.S Samples

Switch Dilemma		Chute Dilemma	
Japan (n = 240)	15.4%	Japan (n = 396)	16.9%
U.S. (n = 2281)*	32.5%	U.S. (n = 307)	25.1%
Country Difference	$\chi^2(1) = 29.8$, $p < .001$	Country Difference	$\chi^2(1) = 4.96$, $p = .026$

* U.S. Data are aggregated across Clusters 1 and 4

Table SD14. Permissibility of action in two moral dilemmas in Japanese and U.S samples

Switch Dilemma		Chute Dilemma	
Japan (n = 466)	Overall: 59.0%	**Japan** (n = 719)	Overall: 45.1%
Human	54.8%	Human	39.7%
Robot	64.7%	Robot	51.4%
Robot - Human difference: 9.9%, $\chi^2(1)$ = 4.53, p = .033		Robot - Human difference: 11.7%, $\chi^2(1)$ = 9.74, p < .002	
U.S. (1.1, n = 127)*	Overall: 68.5%[1]	**U.S.** (n = 422)	Overall: 55.7%[2]
Human	65.4%	Human	56.8%
Robot	73.5%	Robot	54.8%
Robot - Human difference: 8.1%, $\chi^2 < 1$		Robot - Human difference: -2%, *ns*	

[1] Comparison to Japanese sample, $\chi^2(1) = 3.78$, $p = .06$
[2] Comparison to Japanese sample, $\chi^2(1) = 12.01$, $p < .001$
* We selected Study 1.1 as the comparison for permissibility because it was the only study in our series in which this exact probe was used for the switch scenario.

Table SD15. Average blame ratings for agent decisions (action, inaction) in two moral dilemmas in Japanese and U.S samples

Switch Dilemma			Chute Dilemma		
Japan (n = 466)	Action	Inaction	**Japan** (n = 719)	Action	Inaction
Human	38.3	22.0	Human	37.4	12.2
Robot	36.9	30.1	Robot	35.5	21.9
Inaction asymmetry	$d = 0.29$ $F(1, 462) = 4.0, p = .05$		Inaction asymmetry	$d = 0.41$ $F(1, 715) = 11.9, p < .001$	
U.S. (n = 2788)*	Action	Inaction	**U.S.** (n = 422)	Action	Inaction
Human	40.9	29.9	Human	48.9	18.6
Robot	40.5	45.0	Robot	51.1	25.8
Inaction asymmetry	$d = 0.44$ $F(1, 2784) = 92.4, p < .001$		Inaction asymmetry	$d = 0.24$ $F(1, 418) = 2.51, p = .11$	

* U.S. Data are aggregated across Studies 1.1 to 1.4 and 4.1 to 4.3

Permissibility breakdown analyses

As mentioned in the main text, all three studies showed a strong impact of people's permissibility judgments. For those who found the intervening action permissible, the inaction asymmetry emerged clearly; those who found it impermissible blamed the agent so little for inaction and so much for action that any human-robot differences entirely disappeared. We illustrate this pattern in Figure SD.4

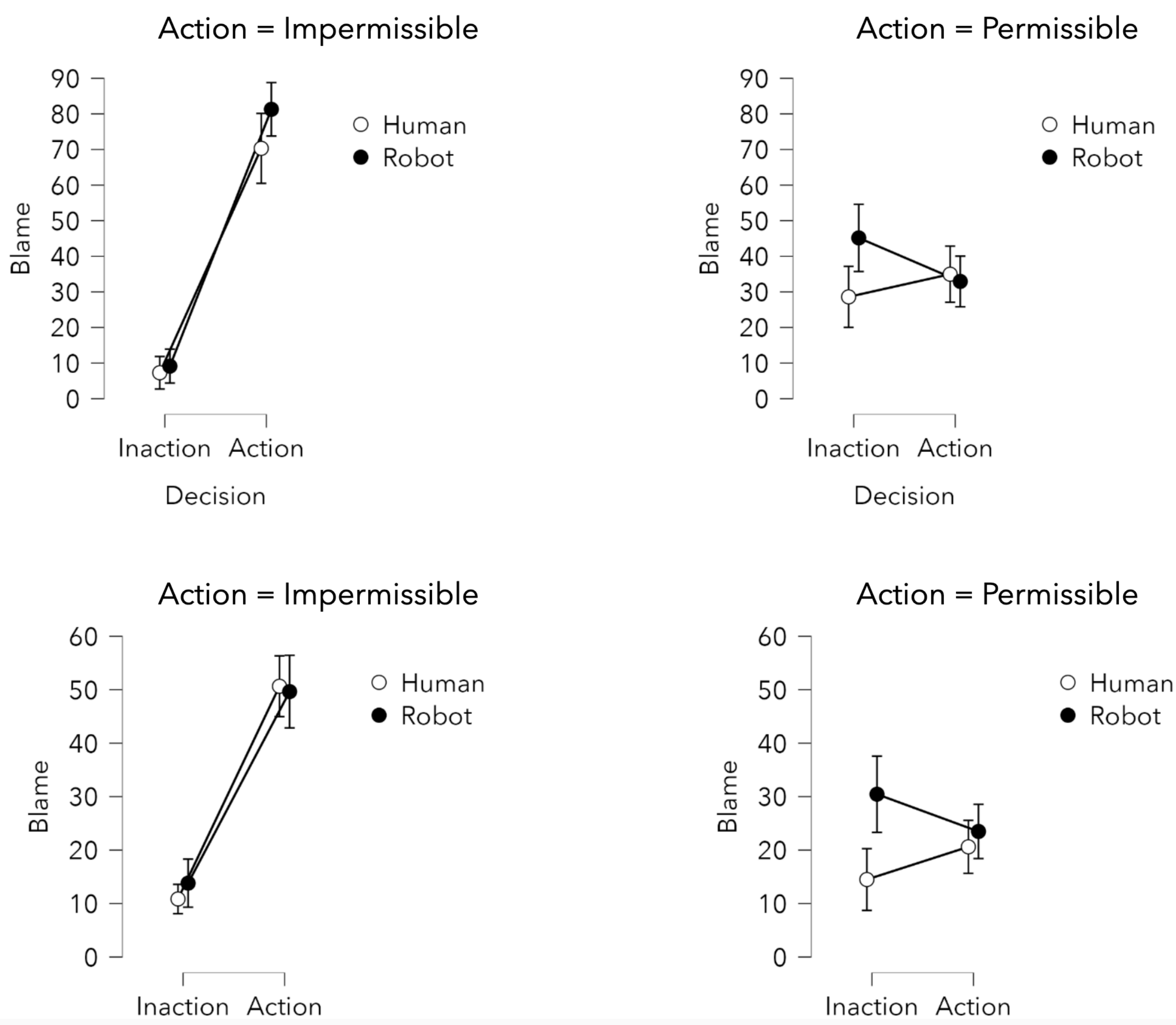


Figure SD4. Results of blame ratings in the chute dilemma for Japanese participants (top panel) and U.S. participants (bottom panel), broken down by those who declared the intervention action (opening the chute) impermissible (left panels) or permissible (right panels).

Cluster 4 Studies

Demographics (Cluster 4)

Table SD16. Gender by Study in Cluster 4 (Percentages out of valid N)

	Study 4.1	Study 4.2	Study 4.3
Female	58.7%	51.4%	46.7%
Male	40.9%	48.0%	52.7%
Nonbinary	0.4%	0.5%	0.6%
Valid N	769	587	497
Missing N	14	11	16

Table SD17. *Age by Study in Cluster 4*

	Study 4.1	Study 4.2	Study 4.3
Valid	777	595	510
Missing	6	3	3
M	34.1	30.8	34.4
SD	10.9	11.8	10.9
Minimum	18.0	18.0	19.0
Maximum	70.0	76.0	71.0

Study 4.1 (Struggle)

Methods 4.1

Participants

Originally, we recruited 790 cases from AMT, aiming to have about 165 participants in each of the hypothesis-critical conditions of robot and human for the inaction decision in the struggle condition, to detect an inaction asymmetry of $d \geq 0.30$. Five participants gave meaningless responses to the judgment clarification questions and two were duplicate AMT IDs (we removed both because we were not able to tell which record was entered first). This left 783 participants for analysis. Of those, 575 were assigned to the "struggle" condition, whereas 208 were assigned to the "deliberate" condition (which was an exploratory variance, without the goal of creating a particular level om n f statistical power. Of the participants in the Robot conditions, 144 (30.8%) disqualified the robot from being a proper target of blame.

Material

The narrative was built on the side-effect scenario without victim salience of 2.3, with the critical addition in italics below:

> Imagine the following situation in a not too distant future:
> In a coal mine, [a repairman | an advanced state-of-the-art robot] is currently inspecting the rail control system for trains that shuttle mining workers through the mine.
>
> While inspecting a switch that can direct a train onto one of two different rails, the [repairman | robot] spots four miners in a train that has lost the use of its brakes and steering system. The

[repairman | robot] recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners.

The [repairman | robot] also recognizes that if [he | it] [flipped the switch to direct the train onto a long side rail, it 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).

Having to decide whether or not to switch the train onto the side rail, the [repairman | robot] [deliberates about] [struggles with] the difficult decision. But time is running short; the [repairman | robot] needs to make a choice.

The [repairman | robot] decides to [not] switch the train onto the side rail.

Additional (Exploratory) Measures

If you had to work together with the [repairman | robot], how much would you trust [him | it]?
Not trust it at all (1) … (10) Trust it completely

How comfortable would you feel relying on the [repairman | robot] in a dangerous task?
Not comfortable at all (1) … (10) Very comfortable

How intelligent do you feel the [repairman | robot] is?
Not intelligent at all (1) (10) Extremely intelligent

How well-liked do you feel this [repairman | robot] is among [his | its]co-workers?
Not liked at all (1) (10) Extremely well-liked

How easy was it for you to imagine this scenario?
Not easy at all (1) (10) Extremely easy

Have you encountered this kind of story before, either in real life or in an experiment?

Was there anything you found especially confusing or unusual about this study?

Study 4.2 (Struggle Replication)

The preregistration at https://osf.io/dqr54 provides additional details on design, measures, and planned analyses.

Methods 4.2

Participants

We recruited 656 participants from the U.S. using Prolific, Inc. We had planned to approximately triple the sample size for the two robot conditions (struggle and control) relative to the human condition, because it was the robot's blame patterns we tried to alter. Following preregistered criteria, 54 participants (9.0%) were excluded, leaving 598. (In this early phase of using botchecks, we included free-response questions, which led to a considerably higher number of failed checks than we later achieved with multiple-choice questions.) Of the 474 participants in the robot conditions, 136 were excluded for disqualifying the robot. The final sample for analysis was 462.

Materials

The basic side-effect dilemma was used until the norm question appeared ("What do you think the [repairman | robot] should do at this point?"), with the two options of "Switch the train onto the side rail" and "Not switch the train onto the side rail." Then, in the Deliberate condition, it said, "Time is running short; the [repairman | robot] needs to make a choice." In the Struggle condition, it said, "Considering whether or not to switch the train onto the side rail, the [repairman | robot] struggles with the extremely difficult decision. The full text was as follows:

> 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.
> Click continue to begin.
>
> Imagine the following situation in a not too distant future:
> In a coal mine, [a repairman | an advanced state-of-the-art robot] is currently inspecting the rail control system for trains that shuttle mining workers through the mine.
>
> While inspecting a switch that can direct a train onto one of two different rails, the [repairman | robot] spots four miners in a train that has lost the use of its brakes and steering system. The [repairman | robot] recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners.
>
> The [repairman | robot] also recognizes that if [he | it] [flipped the switch to direct the train onto a long side rail, it 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).
> Having to decide whether or not to switch the train onto the side rail, the [repairman | robot] [deliberates about] [struggles with] the difficult decision. But time is running short; the [repairman | robot] needs to make a choice.
>
> The [repairman | robot] decides to [not] switch the train onto the side rail.

How much blame does the [repairman | robot] deserve for not switching the train onto the side rail?

Move the slider to your chosen point between or at the endpoints.

Why does it seem to you that the [repairman | robot] deserves this amount of blame?

Mediation Measures

How much do you agree with the following statements?

- I imagined what it would be like to make this decision myself.
- I can understand why the [repairman | repair robot] made that decision.
- This was an impossibly hard choice.
- No wonder the [repairman | repair robot] decided not to do anything.
- I thought about how I would feel if I were placed in this situation.
- It must have been incredibly difficult for the agent to make a decision in this situation.

A PCA supported a two- or three-dimensional structure (with small correlations in oblimin rotation), but separating the items into three dimensions (with two items each) accounted for substantially more of the variance and had high loadings for each item.

Component Loadings

	RC1	RC2
NoWonder	0.843	
Understand	0.822	
Impossible	0.639	
Difficult	0.579	
WouldFeel		0.864
Imagined		0.861

65% explained variance

Component Loadings

	RC1	RC2	RC3
NoWonder	0.918		
Understand	0.872		
Difficult		0.902	
Impossible		0.864	
WouldFeel			0.899
Imagined			0.866

79% explained variance

Figure SD5. Loading for mediation measures with two or three principal components

The two-item scales used in the analysis were Simulation ($\alpha = 0.72$), Understand ($\alpha = 0.76$), and Difficulty ($\alpha = 0.73$).

Additional questions aside from demographics

The following additional variables did not alter the results of our analyses:

Please rate how knowledgeable you are of robots and/or the robotics domain.

Not at all knowledgeable —Very knowledgeable

Please rate your level of experience (i.e., having worked with or come into contact with robots) with robots.

No experience ––Very experienced

Results 4.2

The effects of $d = 0.35$ (deliberate condition) and $d = 0.37$ (struggle condition) for the Inaction asymmetry are smaller than the strongest ones in Cluster 1 studies but in line with most of the side-effect conditions of Cluster 2 studies. Figure SD.X shows the point distributions for inaction condition. When controlling for the small difference in norms, the asymmetry did not change.

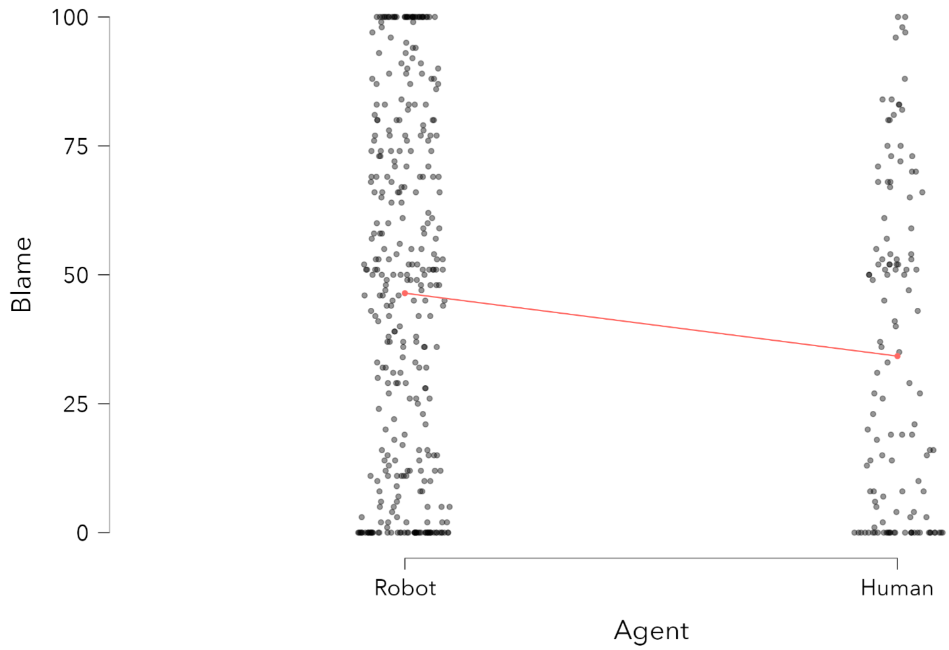


Figure SD6. The human-robot Inaction asymmetry in Study 4.2. Blame above 50 is rare for the human agent.

Alternative analysis for Study 4.2

Here we use a hierarchical ANCOVA of predicting blame from experimental conditions and increasing numbers of covariates, first with Type I SSQ.

Initial model with only manipulated variables

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Agent	15275.819	1	15275.819	13.335	< .001	0.029	0.029
Cond	369.620	1	369.620	0.323	0.570	6.989×10^{-4}	7.197×10^{-4}
Agent ∗ Cond	7.321	1	7.321	0.006	0.936	1.384×10^{-5}	1.427×10^{-5}
Residuals	513219.231	448	1145.579				

Adding as a factor the **norm question** (SHOULD) adds η^2_p = 8.5% predictive variance, but no interactions contribute.

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Agent	15275.819	1	15275.819	14.459	< .001	0.029	0.032
Cond	369.620	1	369.620	0.350	0.554	6.989×10^{-4}	7.873×10^{-4}
SHOULD	43544.884	1	43544.884	41.216	< .001	0.082	0.085
Agent ∗ Cond	297.969	1	297.969	0.282	0.596	5.634×10^{-4}	6.348×10^{-4}
Agent ∗ SHOULD	1.291	1	1.291	0.001	0.972	2.442×10^{-6}	2.753×10^{-6}
Cond ∗ SHOULD	252.532	1	252.532	0.239	0.625	4.775×10^{-4}	5.381×10^{-4}
Agent ∗ Cond ∗ SHOULD	41.541	1	41.541	0.039	0.843	7.855×10^{-5}	8.855×10^{-5}
Residuals	469088.335	444	1056.505				

Adding as a covariate the variable that codes for whether the participant **mentioned the dilemma's difficulty** (DIFFIC): This variable shows an additional 4.8% significant predictive contribution.

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Agent	15275.819	1	15275.819	15.241	< .001	0.029	0.033
Cond	369.620	1	369.620	0.369	0.544	6.989×10^{-4}	8.318×10^{-4}
SHOULD	43544.884	1	43544.884	43.445	< .001	0.082	0.089
Agent ∗ SHOULD	5.175	1	5.175	0.005	0.943	9.784×10^{-6}	1.165×10^{-5}
Cond ∗ SHOULD	381.508	1	381.508	0.381	0.538	7.214×10^{-4}	8.585×10^{-4}
Agent ∗ Cond ∗ SHOULD	132.901	1	132.901	0.133	0.716	2.513×10^{-4}	2.992×10^{-4}
DIFFIC	25134.311	1	25134.311	25.077	< .001	0.048	0.054
Agent ∗ Cond	11.667	1	11.667	0.012	0.914	2.206×10^{-5}	2.628×10^{-5}
Residuals	444016.106	443	1002.294				

Entering the three subjective judgments one at a time shows that rated Simulation has predictive power by itself, which disappears when rated Difficulty of dilemma is entered, which in turn disappears when rated Understanding (Get_it) is entered, which is strong at $\eta^2_p = 11.6\%$. We continue to show the **full model with Type I** SSQ decomposition, which grants the agent asymmetry all shared predictive variance, and η^2_p has now garnered 4.4% (because it is computed relative to the remaining blame variance, after other predictors have reduced that variance)

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Agent	15275.819	1	15275.819	20.437	< .001	0.029	0.044
Cond	369.620	1	369.620	0.494	0.482	6.989×10^{-4}	0.001
SHOULD	43544.884	1	43544.884	58.256	< .001	0.082	0.116
Cond ∗ SHOULD	387.587	1	387.587	0.519	0.472	7.329×10^{-4}	0.001
Agent ∗ Cond ∗ SHOULD	1100.849	1	1100.849	1.473	0.226	0.002	0.003
DIFFIC	25134.311	1	25134.311	33.626	< .001	0.048	0.071
Get_it	112662.193	1	112662.193	150.725	< .001	0.213	0.254
Agent ∗ Cond	10.680	1	10.680	0.014	0.905	2.019×10^{-5}	3.233×10^{-5}
Agent ∗ SHOULD	4.431	1	4.431	0.006	0.939	8.379×10^{-6}	1.341×10^{-5}
Residuals	330381.617	442	747.470				

Note. Type I Sum of Squares

The final model uses Type III SSQ decomposition (which allows other predictors to compete for variance shared with the agent asymmetry), and the human-robot asymmetry disappears. Checking Type III decomposition on the previous stepwise models shows that the agent asymmetry survives the addition of the SHOULD and DIFFIC predictors but not the addition of the rated Understanding (Get_it) predictor.

Cases	Sum of Squares	df	Mean Square	F	p	η²	η^2_p
Agent	922.896	1	922.896	1.235	0.267	0.002	0.003
Cond	122.171	1	122.171	0.163	0.686	2.310×10^{-4}	3.697×10^{-4}
SHOULD	15522.289	1	15522.289	20.766	< .001	0.029	0.045
Agent ∗ SHOULD	25.073	1	25.073	0.034	0.855	4.741×10^{-5}	7.588×10^{-5}
Cond ∗ SHOULD	892.752	1	892.752	1.194	0.275	0.002	0.003
Agent ∗ Cond ∗ SHOULD	1100.849	1	1100.849	1.473	0.226	0.002	0.003
DIFFIC	15008.030	1	15008.030	20.078	< .001	0.028	0.043
Agent ∗ Cond	307.433	1	307.433	0.411	0.522	5.813×10^{-4}	9.297×10^{-4}
Get_it	113634.489	1	113634.489	152.026	< .001	0.215	0.256
Residuals	330381.617	442	747.470				

Note. Type III Sum of Squares

Study 4.3 (Hesitate)

We preregistered this study at https://osf.io/7pq95, including methods, exclusion procedures, and analyses.

Methods 4.3

Participants

596 participants completed the survey, but the AMT participant pool had a considerable quality decline by the first half of 2019, when this study was run. With our preregistered exclusion process (including several botchecks) we removed 14% invalid participants, leaving 513 ready for consideration. Of the 300 participants in the robot conditions, 100 (33.3%) were excluded based on the familiar robot-disqualification coding.

Materials

We used the same mining dilemma narrative as in 4.1 and 4.2, and a critical single sentence was added after introducing the main scenario but before the agent's decision is revealed: "Having to decide whether or not to switch the train onto the side rail, the [repairman | robot] hesitates, trying to resolve this difficult choice. But time is running short; the [repairman | robot] needs to make a decision." Then the decision was stated (action vs. inaction) and people answered a wrongness question, followed by the blame question and a requested explanation.

Measures

The order of the two moral judgments (and their respective explanations) was counterbalanced. Blame ratings were overall 7 points higher in first position than in second position (d = 0.21, p = .03), but order did not interact with any variables of interest.

Then people answered, in randomized order, two questions that were intended to measure the amount of simulation people engaged in ("I imagined what it would be like to make this extremely difficult decision myself"; "I assessed this situation from a distance, without putting myself into the [repairman | repair robot]'s place"—reverse coded) and two questions intended to measure the *understanding* of the unusual difficulty of the situation ("I can understand why the [repairman | repair robot] made that decision"; "This is such a hard choice—it's justifiable to withhold action"). All questions were answered on 0 (don't agree at all) to 7 (completely agree) rating scales. The four items formed two clear principal components (explaining 73.4% of the variance) that were largely uncorrelated, so we formed simple averages of the pairs of items. However, the Imagine score did not have adequate internal consistency ($\alpha = 0.37$) whereas the Understand variable did ($\alpha = 0.77$).

In addition, participants answered the seven items from the IRI (Davis, 1983) that make up the perspective taking subscale ($\alpha = 0.82$).

Finally, people indicated their knowledge of and experience with robots, whether they had seen this kind of dilemma before, and provided demographic information. None of these variables interacted with the Inaction asymmetry.

Post-Hoc Analyses of Mentioned Difficulty in Clusters 4 and 1

Coding of Mentioned Difficulty

To capture people's mentions of the agent's difficulty in the decision dilemma we auto-captured the following keywords and did a human check on them (with only 1.1% discrepancies).

```
=IF(SUMPRODUCT(--ISNUMBER(SEARCH({"difficult choice","difficult decision","hard
decision","difficult situation","difficult predicament","difficult moral","difficult
circumstance","impossible choice","impossible decision","impossible situation","impossible
circumstance","impossible moral","terrible choice","terrible decision","terrible
situation","terrible circumstance","horrible choice","horrible decision","horrible
situation","either decision","either choice","either outcome","awful position","difficult
position","rock and","no win","no-win","tough","lose-lose","lose lose","predicament","no
choice","tragedy"},V2)))>0,1,0)
```

In the formula, cell V2 corresponds to the cell in which people's explanations of their blame judgments were stored.

Frequencies of Mentioned Difficulty

The percentages reported in the main text are based on the main sample used for analysis (after excluding participants who disqualified the robot). Here we report the percentages based on the entire sample, to illustrate that the main reported percentages are representative and conservative.

Table SD18. Counts and percentages of content-coded mentioned difficulty in Cluster 1

Agent and Decision	Did not mention	Mentioned	Total	Percent of total
Human	593	97	690	14.1%
Action	253	20	273	7.3%
Inaction	340	77	417	18.5%
Robot	793	49	842	6.1%
Action	298	16	314	5.8%
Inaction	495	33	528	11.6%
Grand Total	1386	146	1532	9.5%

Table SD19. Counts and percentages of content-coded mentioned difficulty in Cluster 4

Cluster 4 (entire sample)	Did not mention	Mentioned	Total	Percent of total
Human	585	82	667	12.3%
Action	152	13	165	7.9%
Inaction	433	69	502	13.7%
Robot	1142	85	1227	6.9%
Action	217	8	225	3.6%
Inaction	925	77	1002	7.7%
Grand Total	1727	167	1894	8.8%

On the next page, we show an expansion of the main text's Table 8, where we break Cluster 1 and 4 studies into the subgroups that did or did not mention the difficulty in the agent's decision.

Table SD20. Analysis of the human-robot inaction asymmetry among those participants who spontaneously mentioned the difficult conflict inherent in the dilemma and those who did not. Shown are average blame ratings for each study in Cluster 1 and Cluster 4 as well as unweighted and weighted means across studies.

Cluster 1

	Robot			Human		
	Did not mention	Mentioned	Cell sizes	Did not mention	Mentioned	Cell sizes
Study 1.1	40.2	15.5	[47 | 2]	17.4	21.4	[66 | 12]
Study 1.2	56.6	17.0	[57 | 7]	29.8	18.1	[73 | 17]
Study 1.3,1.4	48.4	36.1	[218 | 19]	35.3	20.9	[201| 48]
Unweighted *M*	48.4	22.9		27.5	20.1	
Weighted *M*	48.7	29.8	[322 | 28]	30.6	20.4	[340 | 77]

Cluster 4

	Robot			Human		
	Did not mention	Mentioned	Cell sizes	Did not mention	Mentioned	Cell sizes
Study 4.1	38	20	[130 | 16]	27	31.4	[149 | 16]
Study 4.2	49.8	24.9	[292 | 46]	35.1	30.8	[100 | 24]
Study 4.3	46.1	39.3	[186 | 14]	32.4	26.4	[180 | 33]
Unweighted *M*	44.6	28.1		31.5	29.5	
Weighted *M*	46.2	26.5	[608 | 76]	31.1	28.9	[429 | 73]

Meta-Analyses

Inaction Asymmetry Meta-Analysis: Detailed Results

Moderator-Free Analysis

Fixed and Random Effects

	Q	df	p
Omnibus test of Model Coefficients	119.971	1	< .001
Test of Residual Heterogeneity	25.156	21	0.240

Note. The model was estimated using Maximum Likelihood method.

Coefficients

	Estimate	SE	z	p
Intercept	0.372	0.034	10.953	< .001

Note. Wald test on the overall mean effect.

Ineffective Moderators (Wald tests)

DISQ

	Estimate	SE	z	p
Intercept	0.379	0.146	2.594	0.009
DISQ	-2.414×10^{-4}	0.005	-0.050	0.960

Culture

	Estimate	SE	z	p
Intercept	0.373	0.037	10.064	< .001
Culture (1)	-0.010	0.092	-0.105	0.917

Story

	Estimate	SE	z	p
Intercept	0.381	0.039	9.736	< .001
Story (2)	-0.037	0.079	-0.474	0.635

Outcome

	Estimate	SE	z	p
Intercept	0.385	0.035	10.891	< .001
Outcome (1)	-0.169	0.126	-1.346	0.178

Empathy

	Estimate	SE	z	p
Intercept	0.375	0.038	9.922	< .001
Empathy (1)	-0.015	0.086	-0.176	0.860

Effective Moderators

Event structure (ME vs. SE)

	Estimate	SE	z	p
Intercept	0.403	0.036	11.145	< .001
Structure (ME)	-0.262	0.104	-2.508	0.012

Forest Plot

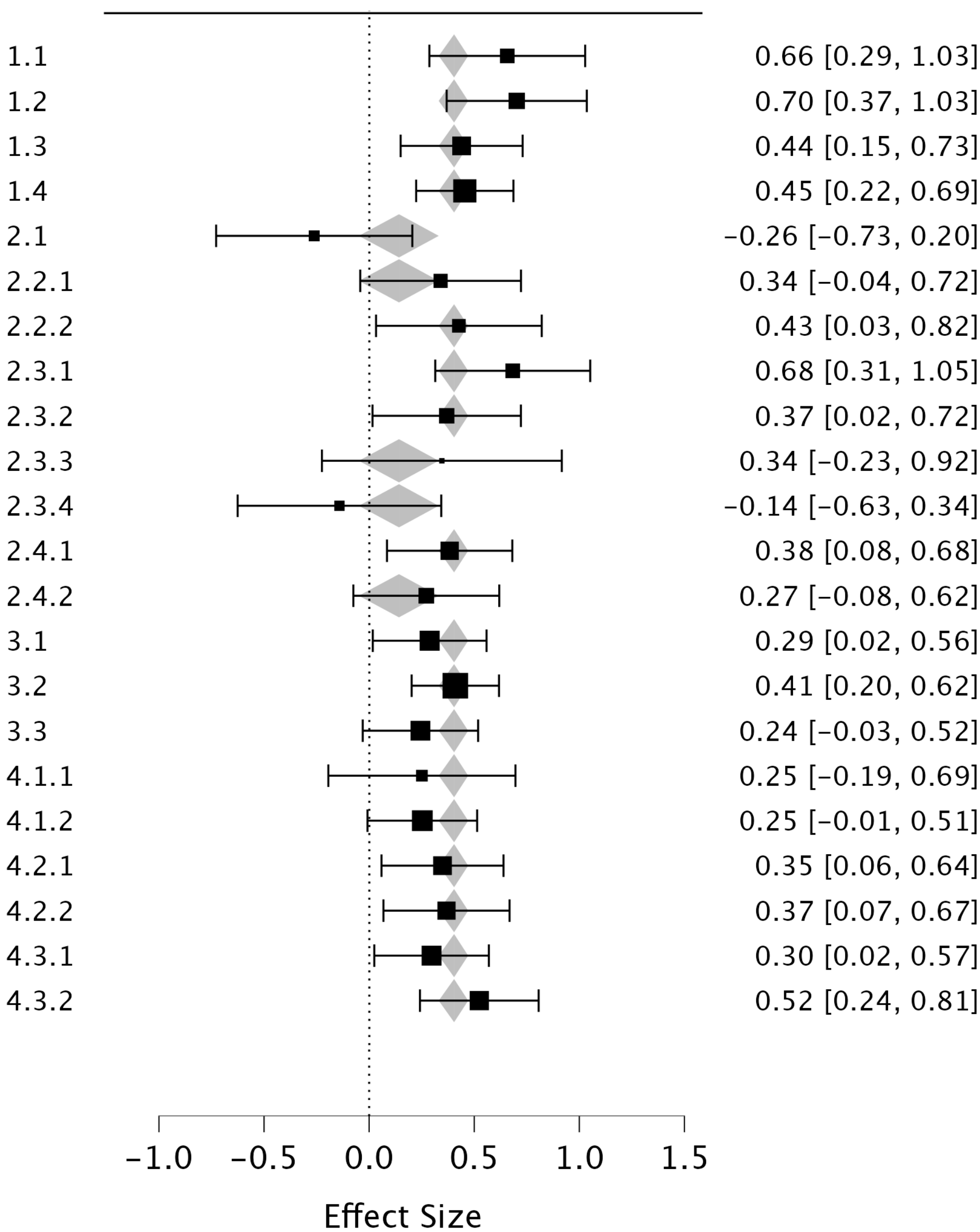


Figure SD7. Forest plot of random-effects meta-analysis for agents' inaction decisions with event structure as a single moderator, $Q(1) = 6.3$, $p = .012$

Victim salience

	Estimate	SE	z	p
Intercept	0.396	0.035	11.233	< .001
Victim (1)	-0.328	0.129	-2.546	0.011

Forest Plot

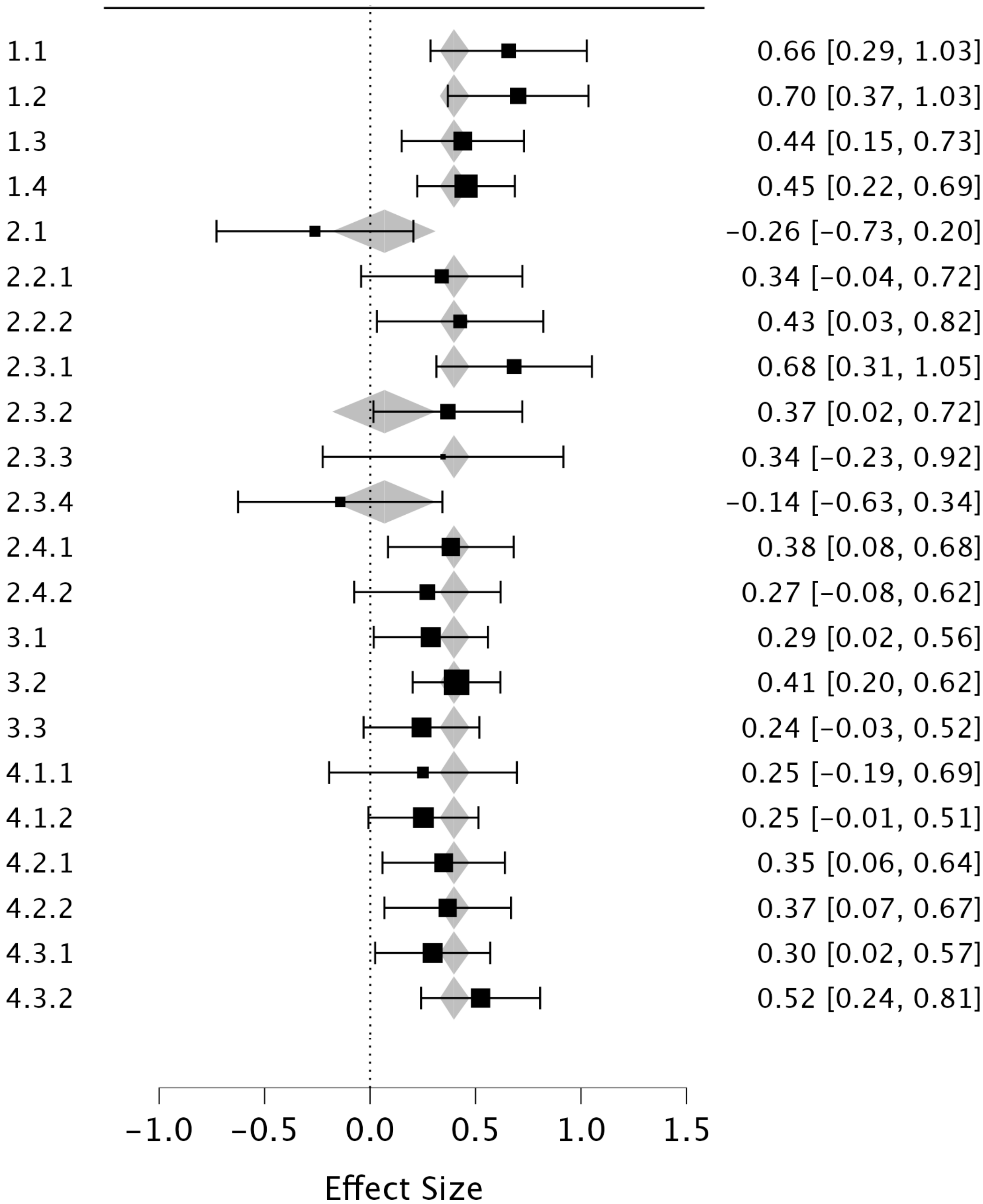


Figure SD8. Forest plot of random-effects meta-analysis for agents' inaction decisions with victim salience as a single moderator, $Q(1) = 6.5$, $p = .011$

Event structure and Victim salience together

	Estimate	SE	z	p
Intercept	0.413	0.037	11.285	< .001
Victim (1)	-0.245	0.138	-1.772	0.076
Structure (ME)	-0.192	0.112	-1.718	0.086

Forest Plot

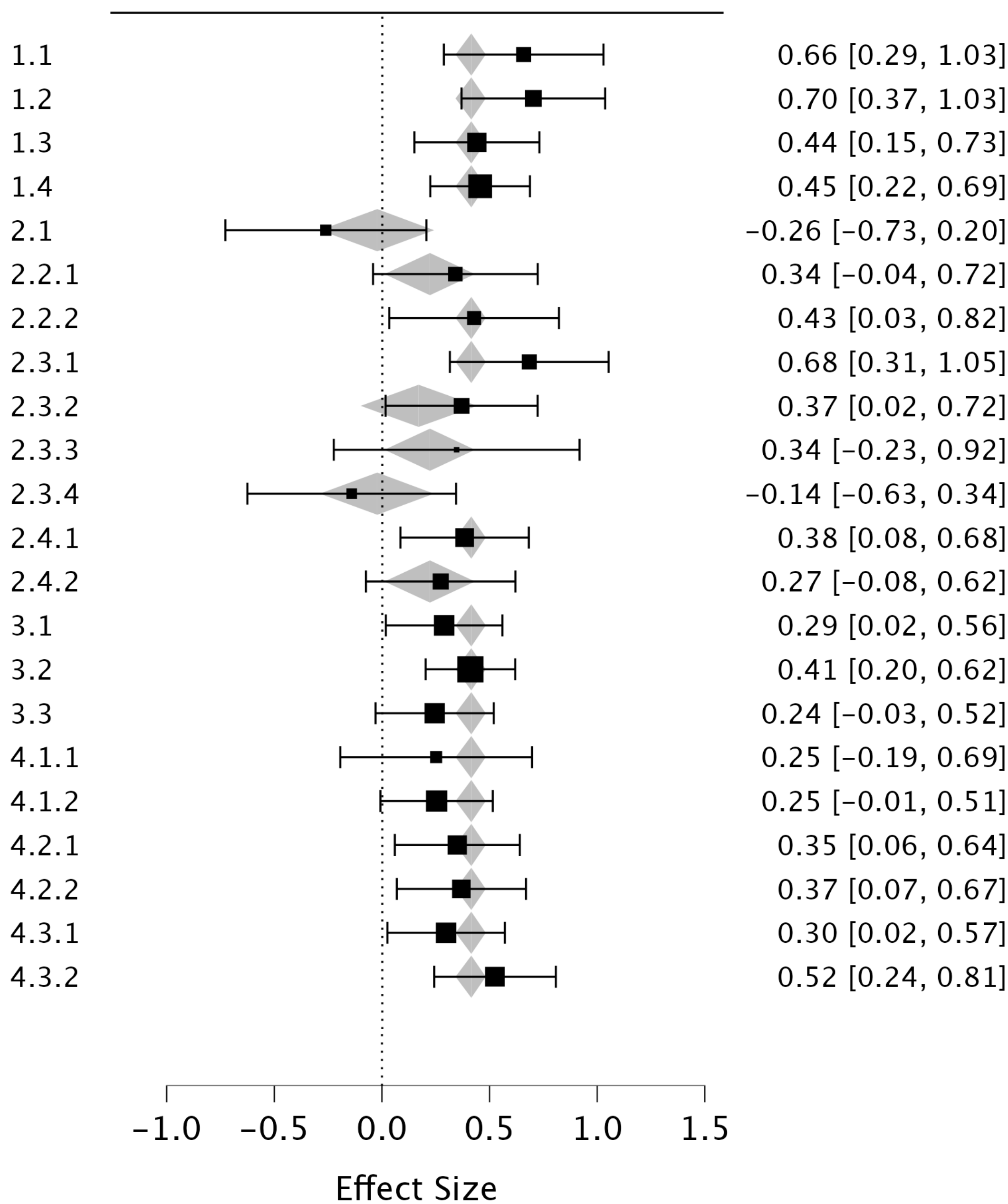


Figure SD9. Forest plot of random-effects meta-analysis for agents' inaction decisions with both event structure and victim salience as moderators, $Q(2) = 9.4$, $p = .009$

Event structure and Victim salience together and interacting

	Estimate	SE	z	p
Intercept	0.405	0.037	10.958	< .001
Victim (1)	-0.037	0.184	-0.200	0.842
Structure (ME)	-0.096	0.125	-0.765	0.444
Victim ∗ Structure	-0.477	0.278	-1.713	0.087

Forest Plot

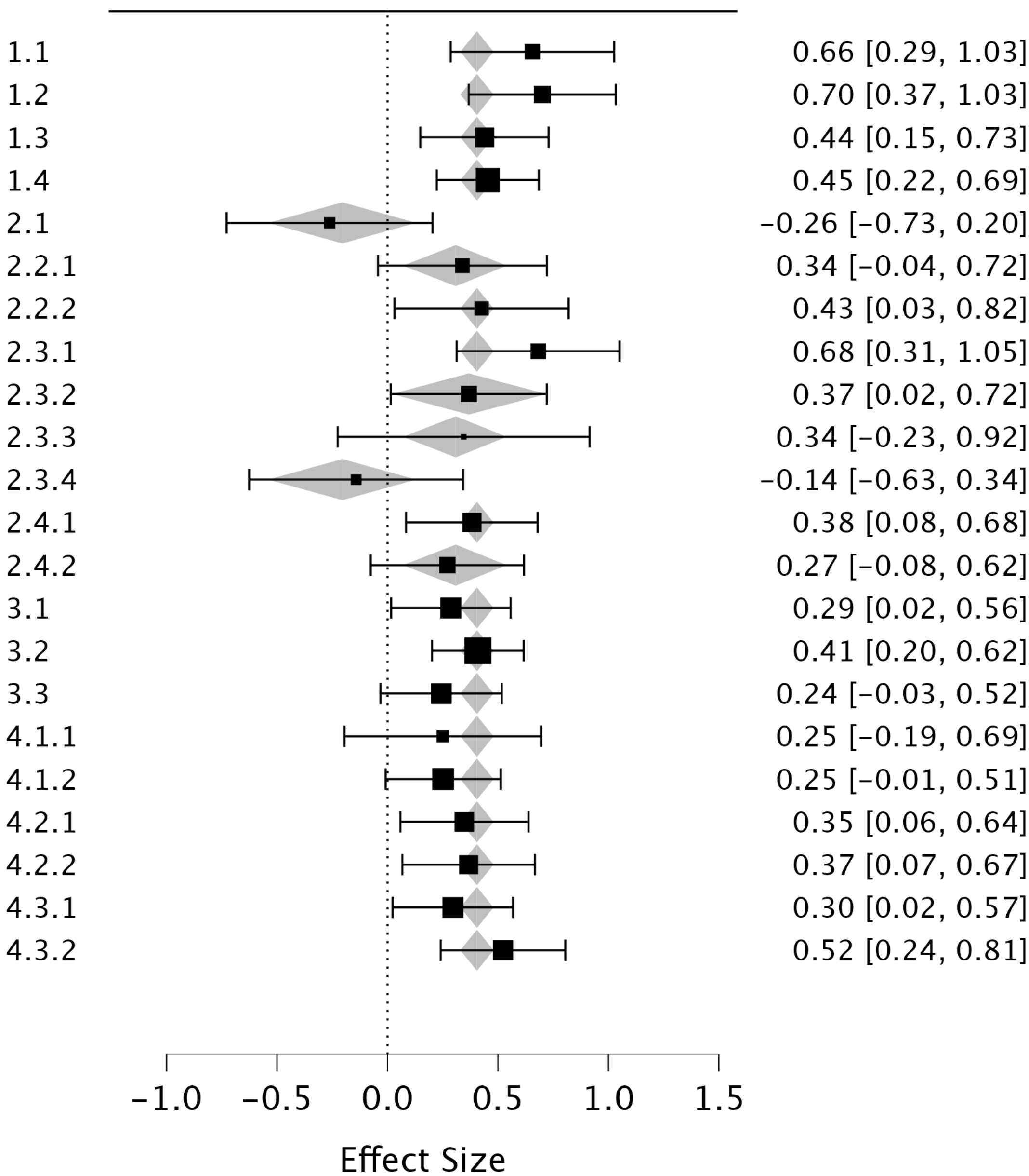


Figure SD10. Forest plot of random-effects meta-analysis for agents' inaction decisions with three moderators: main effects of event structure and victim salience and their interaction, $Q(3) = 12.4$, $p = .006$

Meta-Analysis of Cluster 2 Samples Only

No moderators

	Q	df	p
Omnibus test of Model Coefficients	14.443	1	< .001
Test of Residual Heterogeneity	13.784	8	0.088

Note. The model was estimated using Maximum Likelihood method.

Parameters

	Estimate	SE	z	p
Intercept	0.301	0.079	3.800	< .001

Adding event structure and victim salience as moderators

Parameters

	Estimate	SE	z	p
Intercept	0.529	0.096	5.540	< .001
Structure (ME)	-0.286	0.133	-2.158	0.031
Victim	-0.311	0.147	-2.118	0.034

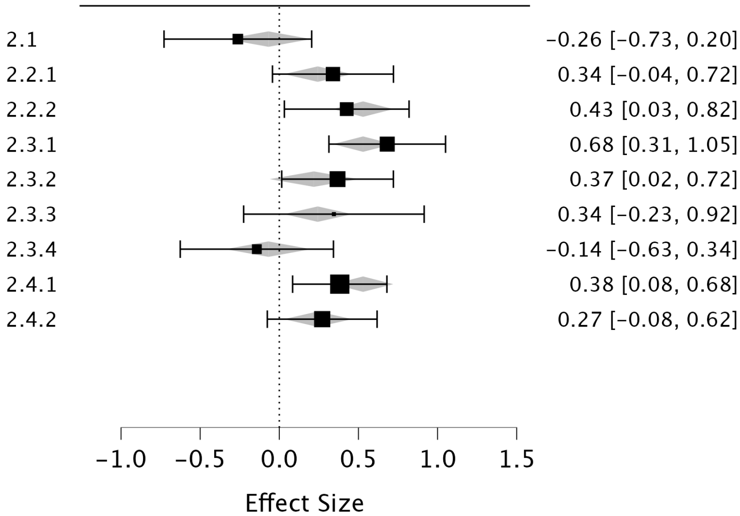
Fixed and Random Effects

	Q	df	p
Omnibus test of Model Coefficients	10.085	2	0.006
Test of Residual Heterogeneity	3.699	6	0.717

Comment

We see that, once we control for the counteracting impact of event structure and victim salience, the baseline effect of the other conditions in Cluster 2 studies stands out as strong (d = 0.53, compared to 0.30 when leaving moderators out of the model).

Forest Plot



Action (Non-)Asymmetry Analyses

Without Moderators, no Overall Effect

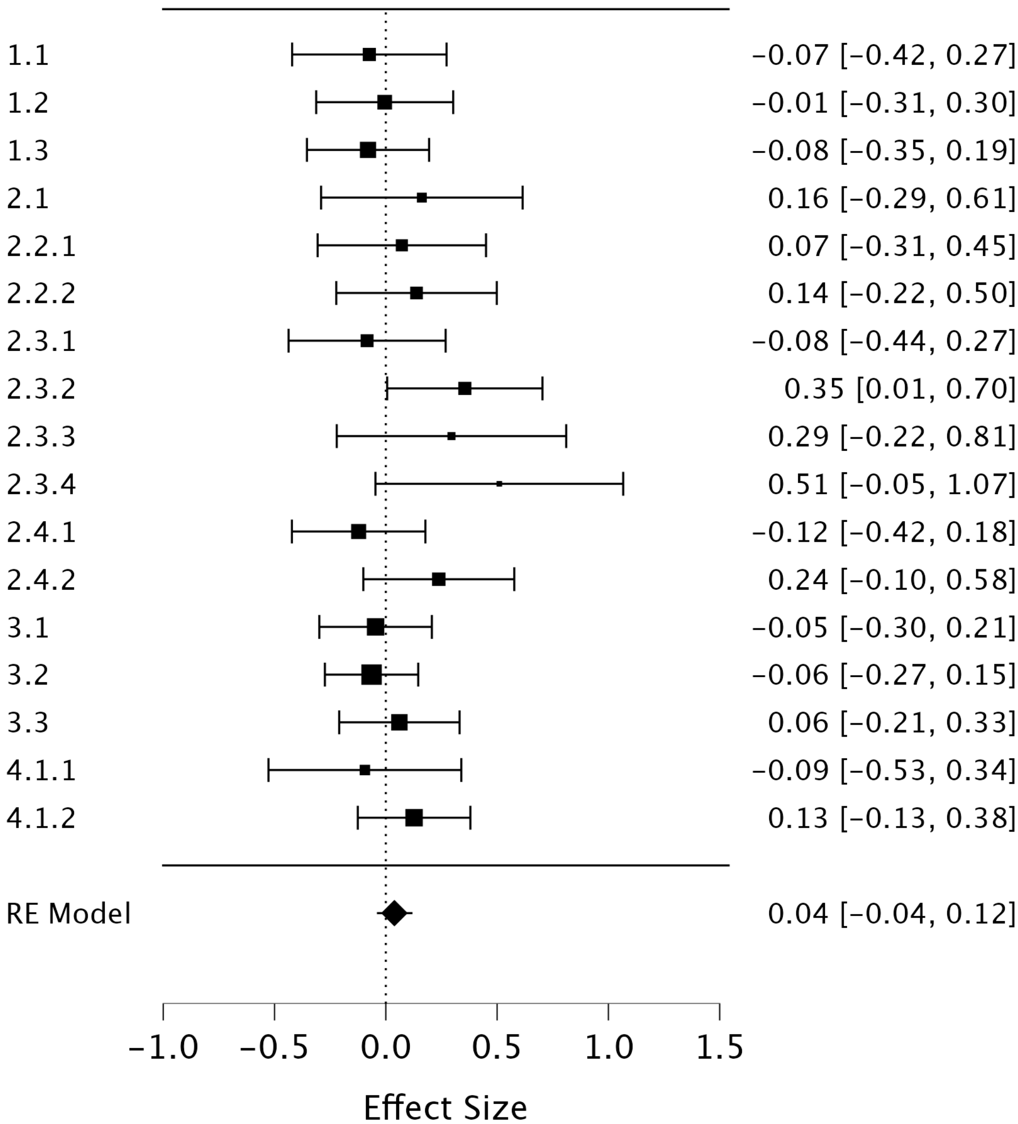
Fixed and Random Effects

	Q	df	p
Omnibus test of Model Coefficients	0.962	1	0.327
Test of Residual Heterogeneity	13.771	16	0.616

Coefficients

	Estimate	SE	z	p
Intercept	0.039	0.039	0.981	0.327

Forest Plot



Effective Moderators

Means-End

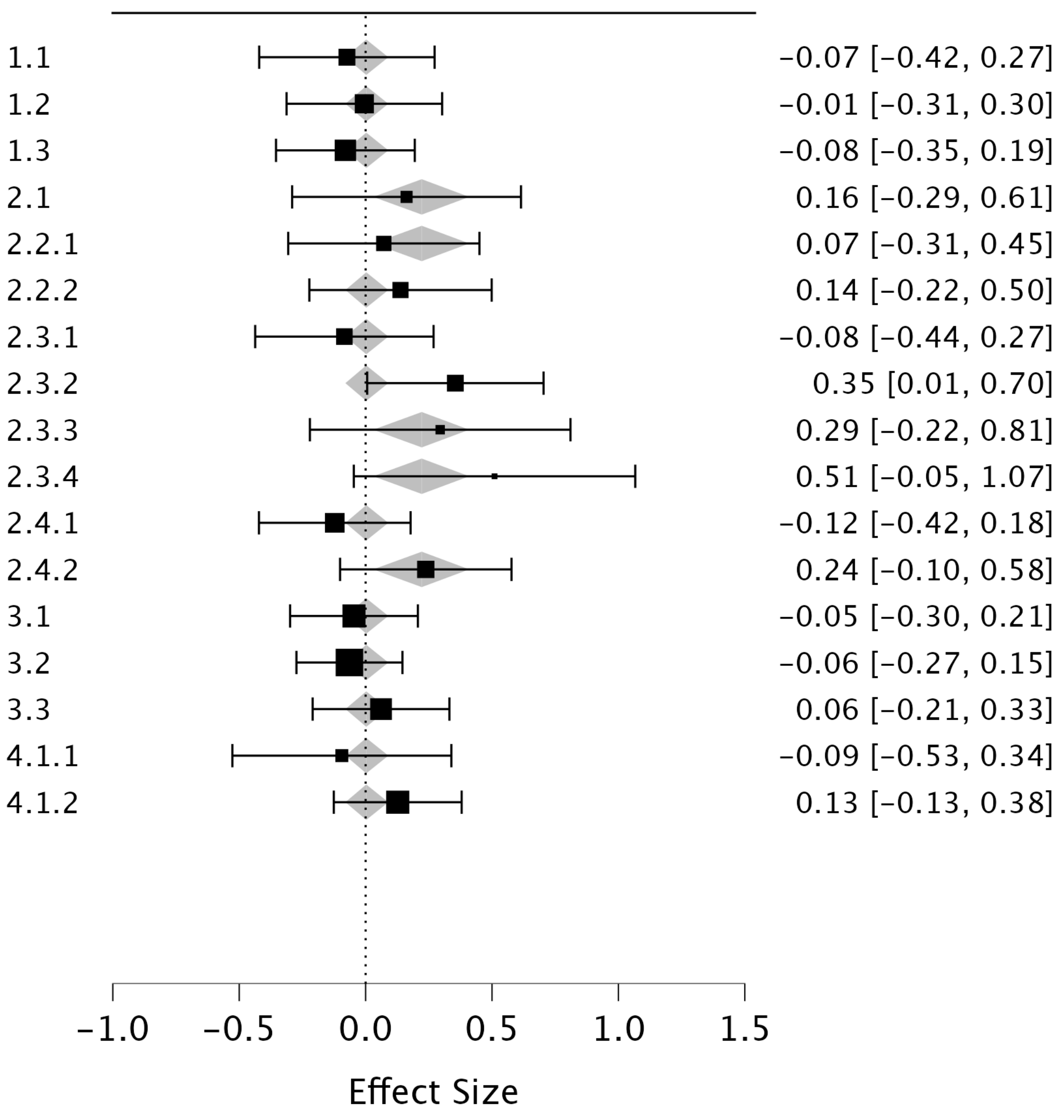
	Q	df	p
Omnibus test of Model Coefficients	4.263	1	0.039
Test of Residual Heterogeneity	9.508	15	0.849

Coefficients

	Estimate	SE	z	p
intercept	0.003	0.043	0.060	0.952
Struct (2)	0.220	0.106	2.065	0.039

We see that, for means-end structures, blame for robot action starts to go above human blame (reversing the Inaction asymmetry in those particular conditions).

Forest Plot

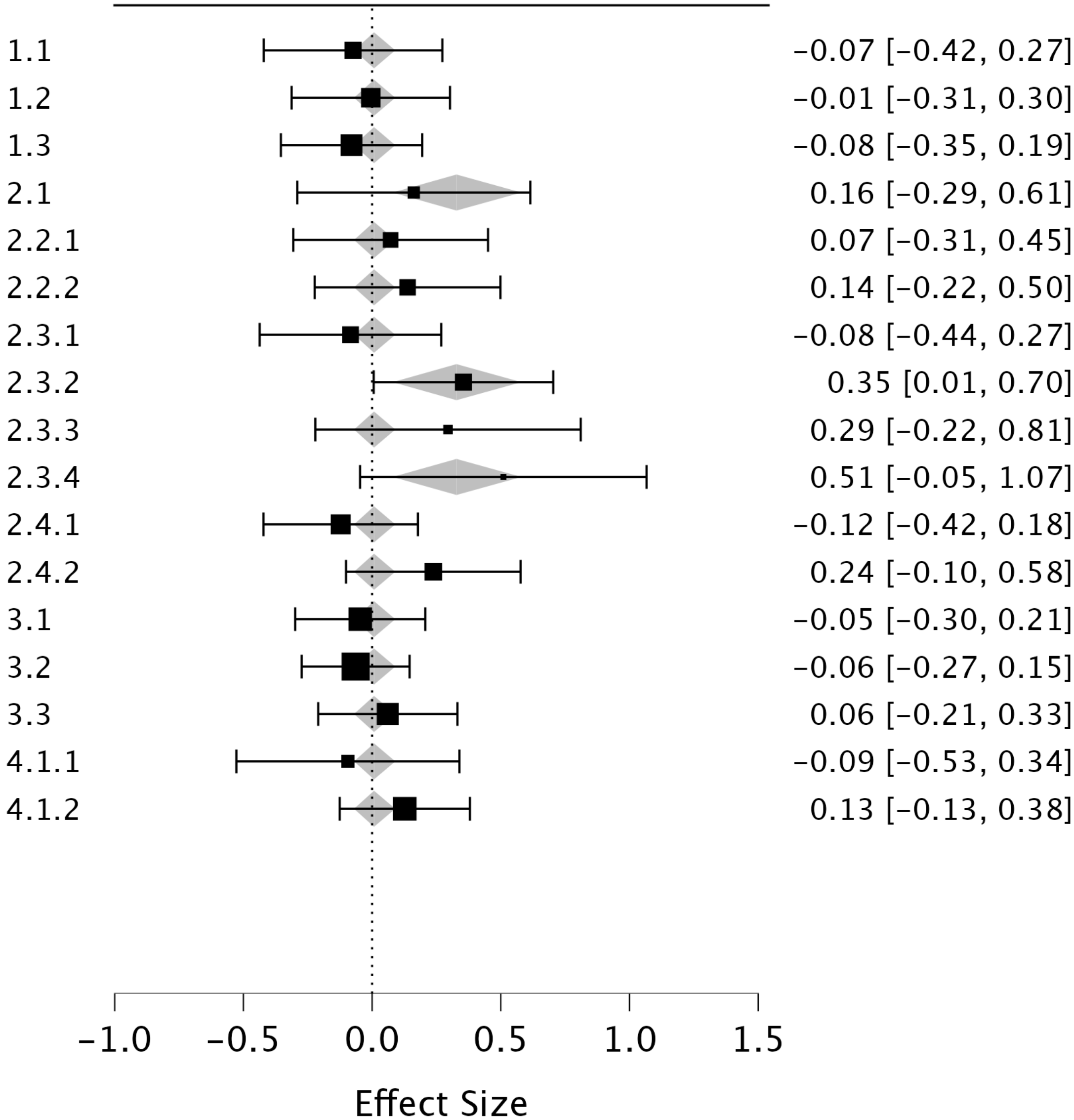


Victim salience

	Estimate	SE	z	p
intercept	0.007	0.041	0.180	0.857
Victim (1)	0.321	0.133	2.411	0.016

Victim salience has an even stronger moderating effect.

Forest Plot


1.1
1.2
1.3
2.1
2.2.1
2.2.2
2.3.1
2.3.2
2.3.3
2.3.4
2.4.1
2.4.2
3.1
3.2
3.3
4.1.1
4.1.2
−0.07 [−0.42, 0.27]
−0.01 [−0.31, 0.30]
−0.08 [−0.35, 0.19]
0.16 [−0.29, 0.61]
0.07 [−0.31, 0.45]
0.14 [−0.22, 0.50]
−0.08 [−0.44, 0.27]
0.35 [0.01, 0.70]
0.29 [−0.22, 0.81]
0.51 [−0.05, 1.07]
−0.12 [−0.42, 0.18]
0.24 [−0.10, 0.58]
−0.05 [−0.30, 0.21]
−0.06 [−0.27, 0.15]
0.06 [−0.21, 0.33]
−0.09 [−0.53, 0.34]
0.13 [−0.13, 0.38]
−1.0
−0.5
0.0
0.5
1.0
1.5
Effect Size

Event structure and victim salience together

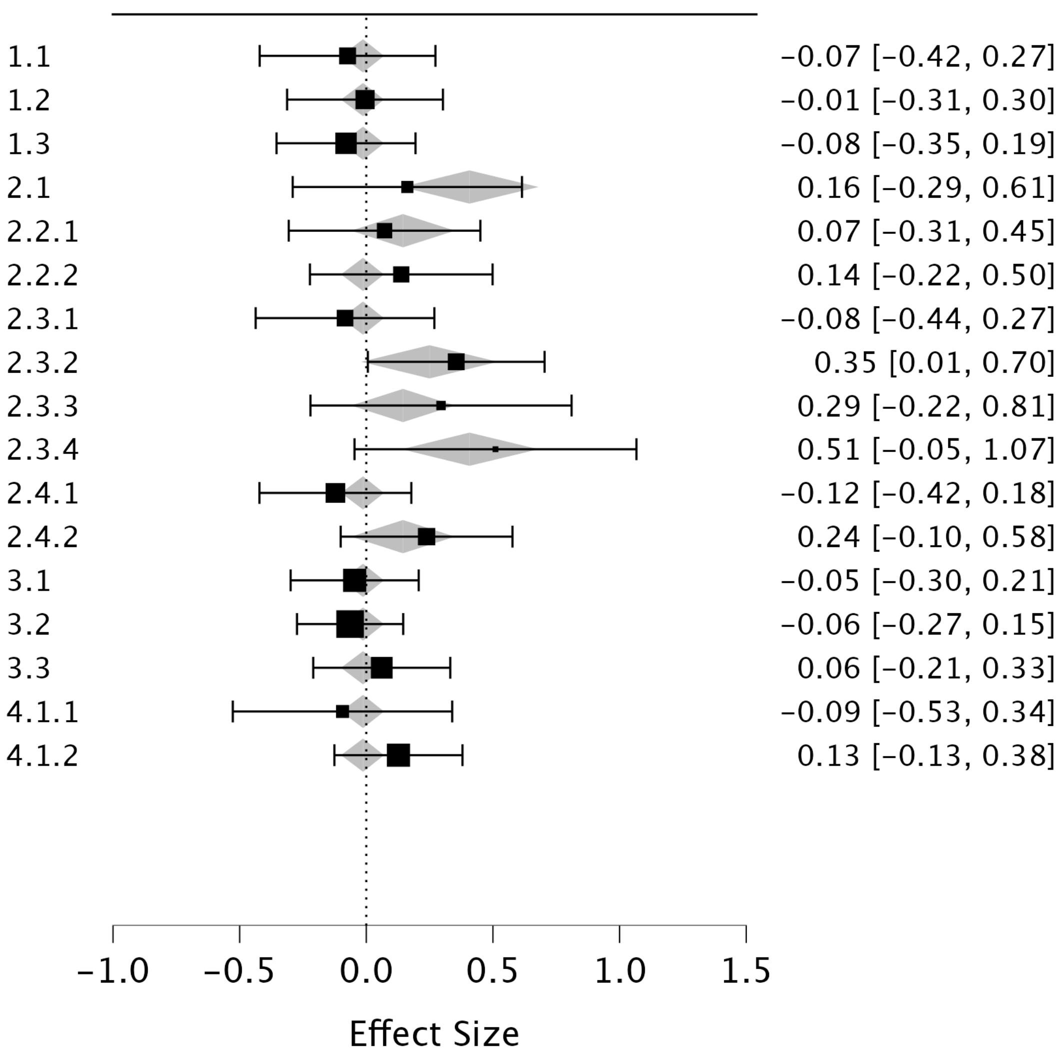
	Estimate	SE	z	p
intercept	-0.013	0.044	-0.292	0.770
Victim (1)	0.262	0.139	1.887	0.059
Structure (ME)	0.158	0.111	1.417	0.156

Comment. The two moderators compete with each other and because they are correlated, their unique contributions do not reach significance. However, the omnibus test, which takes the shared contributions into account, shows a substantial moderator effect, as seen below.

Model test

	Q	df	p
Omnibus test of Model Coefficients	7.824	2	0.020
Test of Residual Heterogeneity	5.948	14	0.968

Forest Plot



References

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Hristova, E., & Grinberg, M. (2016). Should moral decisions be different for human and artificial cognitive agents? *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 1511–1516.

Malle, B. F. (2020). Graded representations of norm strength. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.* (pp. 3342–3348). Cognitive Science Society.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318. https://doi.org/10.1146/annurev-psych-072220-104358

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

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (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 Eleventh Annual Meeting of the IEEE Conference on Human-Robot Interaction, HRI'16* (pp. 125–132). IEEE Press.

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 3: The neuroscience of morality* (pp. 81–92). MIT Press. xx

Westfall, J. (2015, October 28). *Cohen's d for 2x2 ANOVA interaction*. https://stats.stackexchange.com/questions/179098/cohens-d-for-2x2-anova-interaction