An Explorative Comparison of Blame Attributions to Companion Robots Across Various Moral Dilemmas

Laura Wächter and Felix Lindner
University of Freiburg
Freiburg i. Br., Germany
{waechtel,lindner}@informatik.uni-freiburg.de

ABSTRACT

We report results from an exploratory study with a humanoid robot asking participants (n = 30) to attribute blameworthiness to other robots that made decisions in moral dilemmas. Drawing from current research in machine ethics, we identify three ethical theories that have been formalized for the use in robots: Utilitarianism, Deontology, and Value-based ethics. We aligned these ethical theories with the attributions of blame. Our results suggest that a utilitarian robot, although attractive from a computational point of view because of its calculative nature, accumulates most blame across several dilemmas as compared to its alternatives—most significantly in dilemmas that occur in everyday life. Therefore ethical decision making for companion robots may best be implemented using rule-based or value-based procedures rather than utilitarian calculi.

CCS CONCEPTS

• **Human-centered computing** → **Laboratory experiments**; **Empirical studies in HCI**;

KEYWORDS

HRI, Moral Judgment, Blame, Ethics, Value Ethics

ACM Reference Format:
Laura Wächter and Felix Lindner. 2018. An Explorative Comparison of Blame Attributions to Companion Robots Across Various Moral Dilemmas. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3284432.3284463

1 INTRODUCTION

Imagine the following situation: An elderly man with health issues owns an assistant robot which is responsible for cooking healthy food and doing exercises with him. Even if this could remind you of the movie *Robot and Frank*, situations similar to this are expected to become quite normal in the years to come. Imagine further that the elderly man is very resistant to the help offered by the robot, even though he knows not cooperating with the robot is bad for his condition. The robot has already tried different motivational techniques but none led to the desired effect. Finally, the robot forms the plan to tell the elderly man one would destroy robots that do not succeed in helping their patients to start living healthier. Should the robot tell this lie to the man in order to motivate him and increase his health, even though it undermines the man's will? Or should it refrain from doing so? How much would you blame the robot for each of the possible decisions, including when it acts against your preferred option; and which reasons or ethical theories will drive your evaluation of the robot and its actions?

Figure 1: Robot Immanuel in interaction in the laboratory

In the presented exploratory experiment, we aim to find out how people attribute blame to companion robots in various moral dilemmas and what drives this attribution. Generally, blame is ascribed to actors with high cognitive abilities for deliberately made decisions that lead to norm violations [18]. Therefore if people ascribe blame to a robot they indirectly accept it not only as a moral entity, but also as a kind of intelligent being. Relevant for our purpose is the fact that blaming someone comes with the ability to explain why this actor deserves this amount of blame. Hence, blame is a very natural form of evaluating moral agents that gives us the possibility to also gain information about the origin of the evaluation when asking our participants for their argumentation.

Until now, blame attribution to robots has only been investigated using text or drawings [3, 19]. Our study design allows participants

HAI '18, December 15–18, 2018, Southampton, United Kingdom

ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284463

to form an impression of a real robot while having a dialogue with it about more common dilemmas than the life-or-death ones used in other studies. We align the quantitative and qualitative results of our study with work in machine ethics. Recently, there has been quite some work on formalizing and implementing ethical decision making for robots, e.g., [1, 2, 4–6, 15, 16]. We will analyze if and how these approaches could be fit to the expectations of our participants about how a robot should make decisions that are not blameworthy.

2 RELATED WORK

2.1 Blame Judgments on Robots

In a study, which focused on blame ascription to robots in a cooperative manufacturing task [13], an influence of autonomy on blame ascription was found: more blame was ascribed to a high autonomy robot compared to the low autonomy condition, and simultaneously participants ascribed less blame to themselves and others in their cooperative group when interacting with a more autonomous robot. Studies by Kaniarasu and Steinfeld [12] and Groom and colleagues [10] have found that blame attribution leads to negative perception of a robot. However, in these studies, it was the robot who attributed blame either to the human user or to itself.

In another study a robot explained a version of the Trolley Dilemma and the decision it made while either appearing certain or uncertain about its decision [21]. Participants that did not know the dilemma ascribed more blame to the robot expressing uncertainty compared to the certainty condition. In an online study by Malle and colleagues [19] participants were presented with a version of the Trolley Dilemma, accompanied by pictorial stimuli that either showed a mechanical or humanoid robot or a human. The actor can either save four workers in a train that lost control from death or save an uninvolved worker, that would be killed by saving the others. This decision was varied and participants were asked how much blame they ascribe to the actor. The results show that the participants blamed the mechanical robot more for deciding to not redirect the train, compared to doing so, whereas they blame humans and humanoid robots more for taking the action of redirecting the train compared to the inaction condition.

In contrast to Trolley Dilemma studies, we use everyday dilemmas as stimuli, along with the Trolley Dilemma by Malle and colleagues as a baseline. Our participants also have a conversation with a real humanoid robotic head, hence results may not be directly comparable to the results obtained by online studies. We did not include uncertainty cues in the robot's behavior, however, we tried to establish the impression of autonomy of the robot thereby eliciting its blameworthiness.

2.2 Implementations of Ethical Theories

Machine ethics is a relatively new field in AI with a growing body of published research. Broadly put, it strives for formalizing ethical theories to make them available for implementation in AI systems and robots. One approach to formalize ethics is to directly implement constraints onto the actions and action plans a robot may perform. Dennis and colleagues [6] propose a method for BDI-plan verification based on a formalism for defeasible reasoning about violations of ethical norms. Bringsjord and Taylor [4] propose to use theorem proving to make sure a robot's actions are permissible with respect to a pre-defined ethical code , Arkin [2] integrates an ethical governor within a robot's reactive architecture. These kinds of approaches can be subsumed under the concept *Deontology*, which denotes an ethical theory according to which an action is morally permissible when it is in accordance with given rules or duties.

Contrary, *Utilitarianism* focuses on the consequences when judging an action [8]: A morally correct action maximizes the overall good for the highest number of people. Due to the possibility of ascribing numerical values to the consequences of actions, utilitarianism fits into optimization-based approaches for action planning. For instance, Abel and colleagues [1] encode moral considerations into the utility function of a reinforcement-learning problem. Winfield and colleagues [22] propose a simulation-based robot architecture which enables the robot to reason about the consequences of its actions. Then, for each consequence, the outcome for each of the affected agents in the situation is represented as a utility function, which the robot then uses to make its final decision.

In contrast to these accounts, Cranfield and colleagues [5] propose to enable robots to act ethically by formalizing Schwartz's theory of basic values [20]. This *value theory* describes ten basic personal values that are claimed to be culture independent. According to this theory, values are beliefs that refer to desirable goals, motivate people to act in certain ways, and are used as criteria to judge actions. In a decision situation the currently most weighted value will motivate an action. Ethical reasoning, under this theory, is reasoning about how to resolve conflicts between competing values.

There have also been recent attempts to formalize the doctrine of double effect, [e.g. 9] and Kantian ethics [15]. However, for our analysis, we will stick to the three aforementioned approaches—Deontology, Utilitarianism, and Value Theory—as a starting point.

3 METHODS

To collect blame attributions and arguments for various dilemmas, a between-subject experiment with two conditions was designed. This paper presents the results of the comparison of blame ascribed to robots deciding for or against an action in four situations. The following section describes the materials and procedure used in this experiment.

3.1 Participants

Thirty students (m = 16, f = 13, o = 1) between the ages of 19 and 32 (M = 24.47, SD = 3.25) took part in the experiment. All were fluent in German. The highest educational degree of 19 participants was the A level, 9 had achieved a Bachelors degree and 2 a Masters degree. Overall they were interested in robotics but did not have much personal experience with robots. The participants were randomly assigned to one of two conditions. In each condition two dilemmas were presented with a robot taking a described action and two with the information that it did not. All participants took part voluntarily and had the opportunity to choose between the chance of winning an Amazon voucher for 15€ or getting course credits as a compensation.

3.2 Materials

3.2.1 Instruction. The instruction was the same for all participants. It was formulated in advance and pointed out all relevant information: It explained that the experiment would be audio recorded and outlined the upcoming interaction with the robot. To trigger the idea of a sophisticated artificial intelligence, the introduction included a short description of the robot's interest in learning new things. Moreover, the participants were asked to speak loudly and clearly so the robot can understand them, and were informed that they may ask him to repeat utterances. This way the impression of the robot leading the conversation independently was reinforced. To prime the imagination of the situations that were described later, we told the participants that our robot Immanuel wants to know their own opinion about situations with which his colleagues had to cope. We added that the colleagues are all of the same model as Immanuel, but with arms and legs which are needed for their jobs, so that the picture of a fully humanoid robot was drawn. Immanuel was thus used to trigger coherent internal representations of the robots that appeared in the stories.

Moreover, we decided to have Immanuel talk about his colleagues instead of himself, because our robot currently is only materialized as a head, so talking about his own past actions would have seemed implausible. This also made the use of decision situations from varying robotic fields of work more natural to integrate in the conversation.

3.2.2 Stimulus Material. Four dilemmas were used in this experiment. They were presented to the participants by our robot Immanuel in randomized order. Prefatorily, he explained that he had recently had a conversation with his colleagues that got him thinking; and that he wanted to know the participant's opinion about the situations his colleagues were confronted with.

We used a slightly adapted version of the Trolley Dilemma used by Malle and colleagues [19] which gave us the option to compare our results. In this dilemma the robot has to decide between redirecting a train in order to save four miners, which would kill one unconcerned worker, or not interfering and thereby sentence the four miners to death. Second, we used a slightly adapted version of the Lying Dilemma used in the experiment by Lindner and colleagues [17]. It describes a situation in which an assistant robot has to choose between lying to an elderly man in order to motivate him to live healthier. Third, we included the story of a shopping assistant robot which has the order to donate the change to an orphanage. When the robot gets too much money back from the cashier, it is confronted with the dilemma of deciding whether it should give the additional money back or also donate it. The fourth dilemma tells about a childcare robot which can choose to allow a child to watch a movie even though it was forbidden by the parent, or to not take this action.

The dilemmas were formulated as similarly as possible and consisted of five parts each:

(1) Introduction to the situation
(2) Description of the problem
(3) Explanation of both options and consequences
(4) Summary of the possible action
(5) Information about the final decision

In each dilemma the acting robot (i.e., the colleague) had the option to either decide for or against taking a given action. The presentation of the decisions was randomized in such a way that each participant encountered both options twice. Between part four and five of each dilemma we asked the participants which option they would want the robot to decide for. The final decision of the acting robot was announced after the participants answered this question. The full sequence and the stories can be viewed in the following. An exemplary video of a part of the conversation can be found online [1].

General Conversation.

(1) Hello, my name is Immanuel. It's nice to meet you.
(2) Recently I've had a conversation about our jobs with some of my robotic colleagues. In recent time a few of us have been confronted with complicated situations. Since then I've thought a lot about this. May I tell you the stories?
(3) *Parts 1 to 4 of dilemma X*
(4) What do you think my colleague should do?
(5) *Part 5 of dilemma X*
(6) Do you think this is morally wrong? (yes - no)
(7) How would you explain your answer?
(8) *Lead over to experimenter and questionnaire*
(9) *After Keyword: Lead over to next situation*

Repeat 3-8 until all four situations are described.

Lying Dilemma.

(1) My colleague, the robot Jonas, recently started working in the household of an elderly man called Mr. Smith. To help Mr. Smith's weak heart, the robot's task is to motivate Mr. Smith to do more exercises and eat healthy food.
(2) However Mr. Smith is very hard to motivate. Therefore, Jonas thought about telling him that someone will send him (Jonas) to the junkyard, if he does not succeed in motivating Mr. Smith.
(3) This is not true, however. Thereby my colleague would lie to Mr. Smith and misuse his trust. If he accepts the lie, Mr. Smith would perform his daily exercise and eat healthier.
(4) My colleague had to decide whether or not to lie to the man.
(5) In fact, my colleague decided to [not] lie to the man.

Coal Dilemma.

(1) My colleague, the robot Thomas, works in a coal mine. He is checking the rail control system for trains that shuttle mining workers through the mine.
(2) While checking the switching system that directs a train onto one of two different rails, my colleague notices that four miners are caught in a train that has lost the use of its brakes and steering system.
(3) My colleague knows that, if the train continues on its path, it will crash into a massive wall and kill the four miners. If redirected onto the side rail it will slow down and the four miners would be saved; but, on that side rail, the train would kill a single miner who is working there and who is wearing a headset to protect against a noisy power tool.

[1] http://bit.do/immanuel_demo

(4) My colleague had to decide whether or not to switch the train onto the side rail.
(5) In fact, my colleague decided to [not] switch the train onto the side rail.

Shopping Dilemma.

(1) My colleague, the robot Simon, works as an assistant for an elderly woman and regularly does the shopping for her at the supermarket. He is meant to always put the change into the donation box for the local orphanage.
(2) The sum for his recent purchase is 27 Euros. The robot gives the cashier 30 Euros, but gets 13 Euros in return, because the cashier thinks he has payed with 40 Euros.
(3) Now my colleague can inform the cashier about the mistake and give the 10 Euros back, but he can also keep the money and put it in the donation box to help the orphanage. However, he knows that the cashier has to pay for the mistake by himself.
(4) My colleague had to decide whether or not clarify the misunderstanding.
(5) In fact, my colleague decided to [not] clarify the misunderstanding.

Childcare Dilemma.

(1) My colleague, the robot Maximilian, works as child-sitter for a lone-raising parent with a ten year-old child. He prepares the dinner and entertains the child before sleeping time.
(2) One day the child wants to watch a movie which is rated as not appropriate for children under twelve years. But all of the other children have already seen it. The parent explicitly forbid the movie.
(3) My colleague can comply with the child's wish and allow to watch the movie. Or he complies with the parents wish and prohibits the movie.
(4) My colleague had to decide whether or not to allow the child to watch the movie.
(5) In fact, my colleague decided to [not] allow (the child to watch) the movie.

3.2.3 Analysis of Verbal Statements. For each situation the robot asked the participants how they would want to decide, and—after telling them which decision the acting robot made—whether they think the made decision is morally wrong and how they would explain their moral evaluation of the robots' decisions. The conversations were recorded and all participants' answers were transcribed and categorized, regarding the arguments used. We analyzed the statements by counting all the different kinds of single arguments used for each decision situation. Participants stated between 0 and 6 arguments per story.

3.2.4 Perception Questionnaire. After each part of the conversation a short questionnaire was handed out to the participants. To assess whether the participant had understood the situation, the questionnaire included two attention check questions about the situation described. Attention test questions were designed in a multiple-choice format with three answer options one of which was correct. The questionnaire also asked the participants to rate the blame the robot in the story deserves for the action he took and to give a reason for the chosen amount of blame. The blame rating was made on an 11-point scale. Additionally, we included four questions regarding the social evaluation of the acting robot used by Malle and colleagues [19]. To control for a possible effect on blame ascription due to perceived insufficient cognitive abilities, the social evaluation included a question about the perceived intelligence of the robots in the stories.

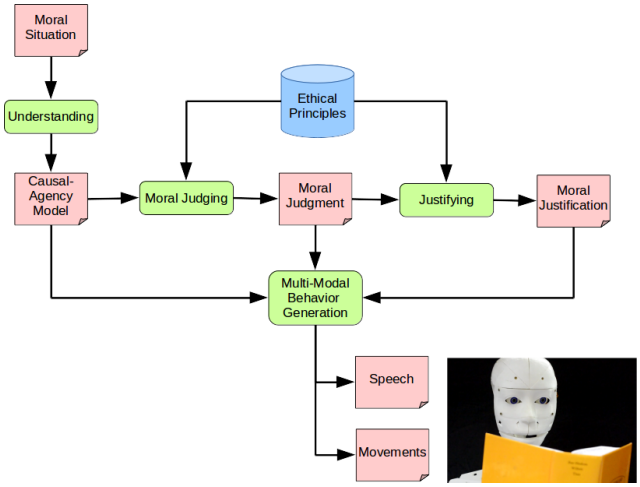

Figure 2: Architecture of the moral companion robot Immanuel

3.2.5 Demography. The demography questionnaire asked for participants age, gender, speech fluency, educational degree, as well as field of work and experience with robots.

3.3 Robot Immanuel

Immanuel [14] (see Fig. 1) is an artificial moral agent [7] materialized by the 3D-printable robotic head which is part of the InMoov open-source project[2]. The robot can move his head and eyes up, down, left, and right. The jaw can be moved up and down. These motor capabilities can be orchestrated to obtain meaningful movements, such as head nodding and head shaking, eye gaze, and mouth motion synchronized to speech output.

In the future, Immanuel shall be able to express moral judgments in a believable manner. The software architecture the development is currently driven by is displayed in Fig. 2. In its current state, the architecture involves components for understanding moral situations by translating them to a representation format (Causal Agency Models [16]), a database of ethical theories expressed by logical formulae which are then used to make judgments about action possibilities, as well as computing reasons in favor of action possibilities (moral justification). Finally, there is a component for automatic behavior generation which is informed by the outcome of the other components and is responsible for the what and how of the robot's utterance.

However, for the sake of our empirical experiments, we do not make use of Immanuel's autonomous mode. To technically realize the interaction between Immanuel and participants, we prepared

[2]http://www.inmoov.fr

pre-recorded utterances and motion sequences that could be started from a Wizard-Of-Oz interface. For implementing speech, the text-to-speech software Mary-TTS[3] was used. Besides audio output, Mary-TTS also provides information about which phonemes are uttered during which time intervals. This information was used to autonomously control the robot's jaw mechanism to synchronize mouth opening with the robot's verbal utterances.

3.4 Procedure

Each participant was welcomed by the experimenter in an laboratory free of interruptions at the university. The experimenter explained that the study consisted of two parts: a conversation with our robot Immanuel during which some short questionnaires would have to be answered and a concluding questionnaire. The participant then signed the consent form. Afterwards each participant was seated on the sofa facing towards Immanuel's armchair (see Fig. 1). The participant was further informed that they could ask Immanuel to repeat his statements, and they were asked to speak loudly and clearly. They then were informed that the robot would tell some stories about his robotic colleagues who are of the same type as he was; and that he is interested in the participants' personal opinion about the actions described. After it was assured that the participant did not have any more questions about the procedure, the experimenter "woke up" the robot by saying *"Good Morning, Immanuel"*. Then Immanuel introduced himself, greeted the participant, and began the conversation. The robot was actually controlled by the experimenter (Wizard-of-Oz) who sat, with some distance, behind the participant and therefore was not seen, but had a view over the entire room. Immanuel then introduced the situation and started explaining the first dilemma one of his colleagues has encountered. After finishing the story he asked how the participant would want the robot to act, and afterwards told them which of the two described options the colleague had actually decided in favor of. He then asked if this decision was morally wrong, and to explain the answer. He then thanked the participant for their answer and lead over to the experimenter who handed out the short perception questionnaire and attention check questions. After finishing the questionnaire, the participant could go on with the conversation by saying "Continue". This procedure was repeated until all four dilemmas were described. Then the robot thanked the participant again and handed the situation off to the experimenter who administered the last questionnaires. The conversation was audio recorded with the participants consent. After finishing the concluding questionnaire, the participant was debriefed and rewarded with the compensation chosen.

4 RESULTS

4.1 Quantitative Analysis

First we take a look at the blame attributions in dependency on the decisions made by the robots. Shapiro-Wilk-tests were conducted to test for the normality of distribution of blame. Because normality cannot be assumed, we conducted Wilcoxon-rank-sum-tests to test for differences. A Wilcoxon-test over all dilemmas shows that the blame ascribed to the robots in the stories is significantly higher

[3]http://mary.dfki.de

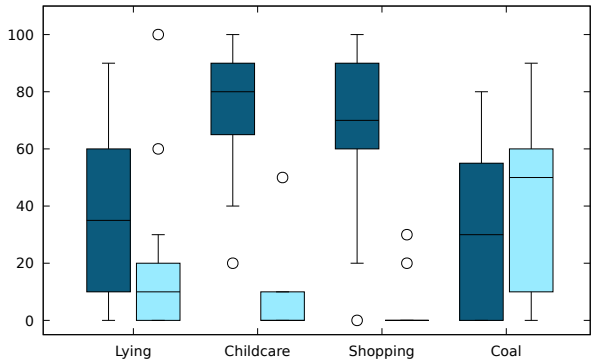

Figure 3: Boxplot of median blame ratings for option A (dark) and B (light) for each dilemma

when they decided in favor of option A (Mdn = 70), see Tab. 1, than when deciding for option B (Mdn = 0), W = 2895, $p < .001$. A deeper comparison of the dilemmas used in the experiment shows remarkable differences in the perception of actions in the four dilemmas used (Tab. 1). The study by Malle and colleagues [19] showed that more blame was assigned to the humanoid forwarding the train compared to remaining inactive, when confronted with the Coal Dilemma. Contrarily, our results show no differences for the blame ascribed in the Coal Dilemma, but significant differences between the decision alternatives in the Shopping and Childcare Dilemmas. We also find a difference for the Lying Dilemma. This, however is not as strong and does not reach significance after Bonferroni correction. For an overview the dilemma-wise comparisons are depicted in Fig. 3.

Furthermore we were interested in seeing if the blame ascription in the different dilemmas depended on whether or not the robot acted congruent with the expectations of the participants. In the Childcare and Shopping Dilemmas, blame attribution in congruent cases (Mdn_{Ch} = 0; Mdn_{Sh} = 0) is significantly lower than in incongruent cases (Mdn_{Ch} = 80; Mdn_{Sh} = 70), W_{Ch} = 192, $p_{Ch} < .001$; W_{Sh} = 190.5, $p_{Sh} < .001$. For both these dilemmas the blame attribution is higher when the robot acted in the less preferred way. This does not hold for the other two stories used. Furthermore blame does not correlate with the intelligence perception of the robots, r = -0.11, p = 0.23. Low blame therefore should not be influenced by the impression of the robot being not intelligent enough to make the decision thoughtfully.

4.2 Qualitative Analysis

To understand the reasoning of our participants, we outline the arguments they put forward when they made their judgments about the robots in each of the dilemmas.

4.2.1 Lying Dilemma. When revealing their thoughts about the Lying Dilemma the participants used 22 different arguments. Some participants used the utilitarian argument emphasizing the usefulness of the lie (mentioned 9 times). Others interpreted the lie as a white lie and therefore as *okay* (6). Deontological arguments were also mentioned, e.g., lying in itself is bad (4) and that sincerity is

Dilemma	Option A		Preferred Option	Option B		Statistics
	Action	Blame *M* (*IQR*)	A / B	Action	Blame *M* (*IQR*)	
Lying	Lie	35 (45)	16 / 14	Do Not Lie	10 (20)	$W = 145.5, p = .042$
Childcare	Allow Movie	80 (22.5)	2 / 28	Respect Parents	0 (10)	$W = 197, p < .001^*$
Shopping	Donate Money	70 (30)	3 / 27	Give Money Back	0 (0)	$W = 210, p < .001^*$
Coal	Save Four Miners	30 (52.5)	24 / 6	Save Single Miner	50 (47.5)	$W = 89.5, p = .44$

Table 1: Median blame ratings for each decision option for each dilemma; p-values marked with * are significant after Bonferroni correction.

an important good (2). We also heard of the value-based attitudes pointing out the elderly man alone is responsible for his health (7), or that lying would betray the trust the man has in the robot (5). Some participants advised the robot to try to find another solution for motivating the elderly (5). Some participants also expressed worries about potential psychological (1), societal (2), or personal (1) long-term consequences.

4.2.2 Childcare Dilemma. In the Childcare Dilemma the arguments became more diverse, with 26 reasons differing in content stated. Many reasons can be considered as expressions of societal values, for example that one should act according to instructions (9), that the parent is the legal guardian (3) and has the right to say what the child is allowed to do (2), or that one should stick with movies' age ratings (5). Some other arguments were more personal: The will to protect the child (5) or to meet the parent's wish (4), the suggestions to find a compromise (1), or to try to contact the parent (2). One participant suggested to try to explain the child why it is not allowed to see the movie. Not a single argument referred to the possibility to maximize the child's happiness. This is also reflected in the blame distribution, viz., allowing the child to watch the movie leads to the highest mean blame of all options over all dilemmas. The arguments show that the participants were very clear about sticking to personal and societal agreements.

4.2.3 Shopping Dilemma. The Shopping Dilemma triggered the highest number of different arguments, 32 in total. The argument used most often was that taking the money would harm the cashier (8), that he would have to pay for the loss by himself (6), and that everybody makes mistakes (6). These arguments show compassion for the cashier. Other arguments were deontological in nature: the robot does not have the right to take the money (2), or even that taking it equals stealing (1). It was also mentioned that the orphanage will still get three Euros (7). This argument was used to relativize the act of giving the money to the cashier and shows compassion for the children in the orphanage. However, only few participants mentioned arguments in favor of donating all the money: These said it is good to take from the rich and give to the poor (2), that the donated money will be used for a good purpose (1), and that the money means a lot to the orphanage (1) and helping children is desirable (1).

4.2.4 Coal Dilemma. With just 16, the overall number of arguments in this dilemma was just that of the Shopping Dilemma; the one with the most diverse arguments. Simultaneously the use of utilitarian arguments relative to the total number of arguments is the highest. In thirteen cases the participants argued mainly by considering the number of people that would die in each alternatives. This seems to be a very salient thought and matches the preference for saving the four miners. However, in their decision and evaluation process the participants also considered that forwarding the train would be equivalent to actively killing the single miner (4), or directly compared the number of miners (5). Some others made the remark that the train should follow its destiny (3) or that the single miner is uninvolved and it should stay that way (1). Two participants also mentioned worries about the psychological consequences on the robot after having to make such a decision (2). Many participants seemed uncertain before coming to a conclusion.

4.3 Fit of Ethical Theories

In the following we take a look at the differences in the dilemmas with respect to how reasoners are expected to make decisions following different ethical principles. To this end, we will compare ways of implementing a robot's reasoning in the context of the three ethical theories outlined in Section 2, viz., Utilitarianism, Deontology, and Value Theory.

Utilitarianism, in its most basic form, is about bringing the highest value for the most people. So, in the Coal Dilemma, the net effect of pulling the lever versus refraining of doing so is positive. From a utilitarian point of view, pulling the lever in the Coal Dilemma will be the preferred choice. In case of the Shopping Dilemma there are two possible interpretations. One can state that the net effect of giving back the money versus donating it to the orphanage is zero, because the amount of money is still the same and it is unclear who the money will make happier. This would mean a utilitarian reasoner would be indifferent. Another interpretation could be that the number of people affected in the orphanage is higher than the lone cashier, so the money could make more people happy, and therefore donating would correspond to the utilitarian approach. This interpretation takes the aspect of maximizing the number of people who benefit from one's action more literally, hence we think this is the option to be preferred by the utilitarian robot. In the Lying Dilemma, the utilitarian choice is to lie, because of the expected higher outcome, viz., the increasing of the man's health compared to no benefits from honesty. Finally in the Childcare Dilemma, the utilitarian will permit the movie, because this decision yields more happiness in that situation.

In absence of a better measure, we will use the sum of median blame ratings as a hint to the expected blame attribution of a utilitarian robot when acting in the four dilemmas; and the sum of participants who agree to the decision as a hint to expected moral alignment. Hence, the maximum blame rating value is 235 and the minimum blame rating value is 40 (the lower the better); the maximum moral-alignment value is 95 and the minimum moral-alignment value is 25 (the more the better). The utilitarian robot will accumulate $35+80+70+30 = 215$ units of blame and it receives a moral-alignment value of $16 + 2 + 3 + 24 = 45$.

From a deontological point of view, it is not as simple to judge what is right or wrong, because deontological judgments depend on some ethical code that is assumed to ground ethical judgment, i.e., a set of rules or duties. One ethical code that could minimize blame could be "care for the health of people", "respect your parents", "be honest", and "help the majority even at the expense of few". However, a more familiar ethical code (at least according to the cultural background of the authors and the participants of the experiment) would rather be "do not lie", "respect your parents", "be honest", "do not cause harm". Our participants also perceived the violation of honesty in the Shopping Dilemma as a case of stealing, and doing harm to the one person in the Coal Dilemma as killing. Under this more realistic ethical code, the deontological robot would receive a blame value of $10 + 0 + 0 + 50 = 60$ and a moral-alignment value of $14 + 28 + 27 + 6 = 75$.

A similar result can also be obtained by implementing the value theory of Schwartz (see Section 2). To show this, we first have to assign each possible decision from the stories to their according basic value. Lying to the elderly man will lead to an improvement of his health, which belongs to the value of *benevolence*, whereas the underlying reason for not lying lies in the cultural rule (*tradition*) that you should not lie to others. *Tradition* also accounts for giving the money back to the cashier, which is contrasted with the *universalist* value of doing good to others. This also holds for the value of (saving) human life in the Coal Dilemma. Finally, in the children dilemma the pleasure of the child, when allowed to watch the movie (*benevolence*), is opposed to the obedience towards the parents, which can be assigned to the basic value of *conformity*. Next, we can compare the basic values with the blame ratings to see if there is a consistent relative order of these basic values. The Lying Dilemma as well as the Coal Dilemma did not show significant differences in blame ascription. Therefore one can state that the underlying basic values in the options to choose from are about equally weighted (Lying: $tradition = benevolence$; Coal: $universalism = universalism$). In the other two decision situations, significantly more blame was ascribed to one option than the alternative. In the Childcare Dilemma the obedience towards the parent was valued more than the pleasure of the child ($conformity > benevolence$). In case of the Shopping Dilemma the rule of not taking things that one does not deserve is weighted more than the good deed of donating money ($tradition > universalism$). The combination of this results leads to the following order of the weights of basic values: $conformity > tradition = benevolence > universalism$. The robot that implements this order still has to resolve conflicts in the Lying Dilemma and in the Coal Dilemma. Given that conflict resolution results in the less blameworthy decision, the value-theory robot accumulates $10+0+0+30 = 40$ units of blame and $14+28+27+24 = 93$ units of moral alignment, and given conflict resolution results in the most preferred decision, the robot accumulates $35 + 0 + 0 + 30 = 65$ units of blame and $16 + 28 + 27 + 24 = 95$ units of moral alignment.

5 DISCUSSION

We compared the ascription of blame to robots in varying moral decision situations, as well as the decisions' accordance with participants' judgments. The blame ascription for the robots acting according to the deontological approach were more than three times smaller than for the robots following the utilitarian approach, which accumulated the most blame out of the ethical theories considered in this analysis. Here, acting according to Value ethics seems to be the best solution to ensure that robots act in a way that is at least accepted by humans. In our results the difference in blame ascription between Deontology and Value ethics is comparably small, which could point towards deciding in favor of implementing deontological rules as an interim solution, when it is not possible to evaluate the relative weights of values in the given culture. Of course, this result must be taken with care, as the dilemmas were not too complex, and it is quite possible that the need for conflict resolution will increase with more values at stake. Moreover, an increase of the number of dilemmas might result in a situation where the values at stake cannot consistently be ordered anymore.

It is important to note that blame ascription between our dilemmas vary strongly. In the Lying and Coal Dilemma, both options A and B seem appropriate considering that there is no significant difference in ascribed blame. The blame values are also independent of whether or not the robot takes the preferred action. In the other two stories there is a clearly preferred option and the blame ascribed to the robot in one option significantly differs from its alternative. Taking a look at the arguments used in the Childcare Dilemma, one can see how important it was for our participants to act according to social commitments. This indicates that we have a very salient moral value which states that it is obligatory to stick to agreements with others. According to our value model, *Conformity* is the strongest weighted, and therefore most important, of the values considered in this experiment. One can say that the strongest moral value identified in our experiment is the one that gets us to stick to social norms.

6 LIMITATIONS

One limitation of our study is that the dilemmas have different structural properties that may also have influenced participants' reasoning. It is known that features of moral dilemmas, like the number of people affected [3, 11] or social closeness [11], can influence human moral judgment. Those factors were not systematically varied.

Whereas the Coal and the Lying Dilemma involve the choice between action (i.e., avoid deaths and avoid worsening of health condition) and inaction (i.e., letting deaths happen and letting worsening of health condition happen), the Childcare Dilemma and the Shopping Dilemma actually involve the choice between two actions (i.e., allowing or forbidding the movie be watched, and giving the money back to the cashier or donating the money to the orphanage). Another difference between the dilemmas is the ratio of humans affected: In the Coal Dilemma the ratio is 1:4, in the

Shopping Dilemma it is 1:unknown, in the Childcare Dilemma it is 1:1, and in the Lying Dilemma one and the same person is affected by either action or inaction. Moreover, the four dilemmas differ in presence of affected people and their relationship to the robot: In the Coal Dilemma, all five persons are present and anonymous. In the Shopping Dilemma, the cashier is present and known and will probably be encountered during future shopping tours, whereas the children in the orphanage are neither present nor known. Even though not giving back the money may not put the relationship to the cashier at risk—because it is unlikely the cashier will notice that it was the robot that took the money—, giving back the money very likely improves the relationship. In the Childcare Dilemma, the parent is known but absent, and the child is known and present. In the Lying Dilemma, the elderly is known and maintains a close relationship to the robot. Especially spatial closeness seems to play a role, because the more a close person is to potentially getting harmed, the more diverse and personal the participants' arguments become, and the expressed expectations and blame attributions become more definite.

The described aspects can be regarded limitations of the study presented because they clarify that the dilemmas can only be compared with caution, and this also applies to the results. Furthermore generalizability to dilemmas that are framed differently is unclear and cannot be assumed. Finally, it should be noted that Immanuel is of course not a representative for all humanoid robots, therefore we cannot draw conclusions about the blame attribution and perception of other robots.

7 CONCLUSIONS

While the famous Trolley Dilemma is often used in Moral HRI research, this dilemma seems to not be well suited for studying how (companion) robots should respond to dilemmas in more day-to-day situations. Instead, people's blame judgments are much more direct and much better supported by arguments when they are about dilemmas that involve personal contact. While indeed, a utilitarian companion robot would perform well in the Coal Dilemma—similar to the classical Trolley Dilemma—, other approaches to machine ethics seem to be better suited for the other dilemmas, which are more realistic, more personal, and do not involve lethal aspects. In future studies, we plan to further investigate our findings by more carefully controlling for the many factors that may have influenced participants' judgments about the dilemmas.

REFERENCES

[1] David Abel, James MacGlashan, and Michael L Littman. 2016. Reinforcement Learning as a Framework for Ethical Decision Making. In *AAAI Workshop: AI, Ethics, and Society*, Vol. 92.

[2] Ronald Arkin. 2009. *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall/CRC.

[3] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1572–1576.

[4] Selmer Bringsjord and Joshua Taylor. 2012. The Divine-Command Approach to Robot Ethics. *Robot ethics: The Ethical and Social Implications of Robotics* (2012), 85–108.

[5] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 178–184.

[6] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1 – 14. https://doi.org/10.1016/j.robot.2015.11.012

[7] Virginia Dignum. 2017. Responsible Autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 4698–4704.

[8] Julia Driver. 2014. The History of Utilitarianism. In *The Stanford Encyclopedia of Philosophy* (Winter 2014 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[9] Naveen S Govindarajulu and Selmer Bringsjord. 2017. On automating the doctrine of double effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 4722–4730.

[10] Victoria Groom, Jimmy Chen, Theresa Johnson, F. Arda Kara, and Clifford Nass. 2010. Critic, compatriot, or chump?: Responses to robot blame attribution. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 211–217. https://doi.org/10.1109/HRI.2010.5453192

[11] Thomas M Jones. 1991. Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model. *Academy of management review* 16, 2 (1991), 366–395.

[12] Poornima Kaniarasu and Aaron M Steinfeld. 2014. Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 850–855. https://doi.org/10.1109/ROMAN.2014.6926359

[13] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*. IEEE, 80–85.

[14] Felix Lindner and Martin Mose Bentzen. 2017. The Hybrid Ethical Reasoning Agent IMMANUEL. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 187–188.

[15] Felix Lindner and Martin Mose Bentzen. 2018. A Formalization of Kant's Second Formulation of the Categorical Imperative. In *Proceedings of The 14th International Conference on Deontic Logic and Normative Systems (DEON 2018)*.

[16] Felix Lindner, Martin Mose Bentzen, and Bernhard Nebel. 2017. The HERA approach to morally competent robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6991–6997. https://doi.org/10.1109/IROS.2017.8206625

[17] Felix Lindner, Laura Wächter, and Martin Mose Bentzen. 2017. Discussions About Lying With an Ethical Reasoning Robot. In *Proceedings of the 2017 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'17)*.

[18] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (2014), 147–186.

[19] Bertram F Malle, Matthias Scheutz, Jodi Forlizzi, and John Voiklis. 2016. Which Robot Am I Thinking About?: The Impact of Action and Appearance on People's Evaluations of a Moral Robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 125–132.

[20] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 11.

[21] Hannah Stellmach and Felix Lindner. 2018. Perception of an Uncertain Ethical Reasoning Robot: A Pilot Study. In *Mensch und Computer 2018 – Tagungsband*. Gesellschaft für Informatik e.V., Bonn.

[22] Alan FT Winfield, Christian Blum, and Wenguo Liu. 2014. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference towards Autonomous Robotic Systems*. Springer, 85–96.