# ‘Well, at least it tried‘

## The Role of Intentions and Outcomes in Ethically Evaluating Robots’ Actions


Bachelor Thesis in Artificial Intelligence by

**Daphne Lenders**

s4433556

Supervised by

**Dr. Willem F. G. Haselager[1]**

[1]Donders Institute for Brain, Cognition and Behaviour




Department of Psychology and Artificial Intelligence

Radboud University

Nijmegen


June 2017

# Table of Content

**Abstract:** More and more people start to use robots in domestic settings. The way a robot is used by its owner depends among other factors on the way the owner perceives the robot's behaviour. If a human considers the robot to act in a moral way, it is more likely to be used to its fullest potential, as a robot that is perceived as moral is usually also perceived as trustworthy. Two important aspects in evaluating behaviour as moral or not, are the intentions and outcomes of the behaviour. The main goal of this research was to find out how the attributions of intentions behind robots' actions and the outcome of their actions influence the moral evaluation of these actions. Video clips showing a robot appearing to have a good or bad intention followed by an action that leads to a good or bad outcome were used to investigate this. It was found that both good intentions and good outcomes have a positive influence on ethical evaluations, while bad intentions and bad outcomes have a negative influence on these evaluations. Just like when evaluating humans' actions, the influence of intentions on moral evaluations was found to be stronger than the influence of outcomes. This was found for both actions in Care/Harm and Authority/Respect contexts.

## 1. Introduction

Over the last year the number of robots has increased a lot, but it is expected that the number of robots for both personal and professional use will grow even more rapidly in the future (International Foundation of Robotics, 2016). Next to the general increase in robots, it is expected that robots will increasingly engage in more complicated behaviour and decision making, sometimes even in an autonomous way (Allen, Wallach & Smit, 2006). This autonomy is reflected in various aspects of robots' behaviour: they sometimes act in an unpredictable or even unwanted way and they can actively learn from their past experiences instead of relying only on pre-programmed patterns (Haselager, 2005).

With this increasing autonomy it is claimed that the need for morality in robots has risen, too (Allen et al. 2006). It is not only important that robots actually act moral, but that they are also considered as moral by their users. This is due to the fact that the way a robot appears and functions, has impact on how people will use it and build relationships with it (Bartneck & Forlizzi, 2004). If a robot does not appear to be moral, it is less likely to be trusted. As a result of decreased trust, robots might be misused by their owners or not used at all (Parasuraman & Riley, 1997). Finding out what makes a robot appear moral is therefore an essential step in making robots more trustworthy and usable to a wider range of users.

In the act of making moral judgments a variety of factors are involved, which cannot be discussed all in this paper. As philosophers however have suggested, two important aspects in morally evaluating actions are the intentions behind the actions and their outcomes (Richardson, 2014). This research focusses on finding out to how the attribution of intentions behind robots' actions and the outcomes of robots' actions influence the ethical evaluation of

that actions. Video-clips of a robot appearing to display an either good or bad intention, followed by an action that leads to a good or bad outcome will be used to investigate this.

Next to that it will be studied whether the influence of intentions and outcomes on ethical evaluations of actions, differ between different moral contexts. Haidt & Kesebir have suggested that there are five moral contexts which are connected to different moral virtues and emotions (2010). It will be investigated whether the influence of intention and outcomes of actions differ in two of these moral contexts, namely Care/Harm and Authority/Respect.

## 2. Theoretical Framework

Normative ethics is known as the philosophical field which aims to give answer to the question about what actions are morally right and morally wrong (Gert & Gert, 2016).

The discussion concerning this question is already centuries old and different theories and models have emerged from it. Two important though very differing theories are Consequentialism and Deontology.

Consequentialists like Jeremy Bentham or John Stuart Mill claim, that only the outcome of an action should be important when morally evaluating that action (Sinnott-Armstrong, 2015). Thus for a consequentialist an action with a bad intention, would be evaluated as equally right as an action with a good intention, as long as both actions lead to the same outcome.

Deontologists like Immanuel Kant however state, that only the intention behind some action should be considered, when evaluating the action as ethical or not (Allexander & Moore, 2016). Different from consequentialists, they would consider an action that was based on a good intention but caused harm as morally right.

While normative ethics focusses on the question which actions are morally right, descriptive ethics concentrates on how persons actually make moral judgements (Gert & Gert, 2016). In descriptive ethics one could e.g. investigate whether humans are more guided by intentions or outcomes of actions when making moral judgements.

Research has already been done on the extent to which the intention and the outcome of some action of a human matters when giving an ethical evaluation to that action. Among others Saxe and Young have found that a good intention behind some action has a stronger effect on the ethical evaluation of an action, than a good outcome (2008). Thus even if the outcome of some action is not good, this action will be evaluated as more ethical than an intentionally bad action that leads to a good or neutral outcome (Saxe & Young, 2008). The

same research however also showed that a positive/neutral outcome of an action has a positive effect on the ethical evaluation of that action: Regardless whether an action was based on a good or bad intention, it is perceived as more ethical when it results into a good outcome rather than a bad one (Saxe & Young, 2008).

Different research like the one of Cushman et al. (2013) and Charness and Levine (2007) have replicated these results. It therefore can be concluded that humans use both deontologist and consequentialists principles when making moral judgments. Seeing however that 'good-intention/bad-outcome' actions are evaluated as more morally right than 'bad-intention/good-outcome' actions, deontologist principles seem to have a higher influence on ethical judgements.

While research like the one of Saxe and Young has focussed on the ethical evaluation of human actions, little research has been done on the ethical evaluation of robot actions. Yet as it has been said before, this type of research is important when trying to make robots appear both moral and trustworthy.

One might argue that it is unconventional to ascribe concepts like 'intention' or 'moral behaviour' to inanimate objects such as robots. This is however where anthropomorphism starts to play a role. Anthropomorphism is described as the tendency of humans to ascribe human features and properties to non-lifelike objects (Fink, 2012). One of its assumptions is that humans prefer to interact with robots as they would with other humans, even though they know that robots are not actually alive (Fink, 2012). Research has suggested that anthropomorphism can be elicited among others by the appearance of the robot and the social cues it gives (Schmitz, 2011). Speech, gestures and other life-like cues can cause a human to ascribe human characteristics, such as moral thoughts and behaviour, to a robot (Schmitz, 2011). This should make it possible to investigate the extent to which the perceived intentions and outcomes of a robot action influence the ethical evaluation of that action.

'Ethical Evaluation' is however quite a broad term, as some psychologists suggest that there may be different moral contexts involved in morally evaluating different actions (Keil, 2014). Two of these psychologists are Haidt and Kesebir, who suggest that there are five different moral domains that are each connected to different virtues and different moral emotions (Keil, 2014). These domains are: 'Care/Harm', 'Fairness/Reciprocity', 'Ingroup/Loyalty', 'Authority/Respect' and 'Purity/Sanctity' (Haidt & Kesebir, 2010).

Due to time constraints in this research, it was chosen to only focus on the ethical evaluation of robot actions within two of these moral contexts. The first context that was

chosen was Care/Harm, which is about protecting and caring for other persons that need help (Haidt & Kesebir, 2010). With the large rise in health care robots over the past year, this context is definitely relevant for the moral evaluation of robots (Salem & Dautenhahn, 2015).

Next to that, the context Authority/Respect was chosen, which is about being obedient to persons with high authority. Since in a human-robot relationship the robot is usually submissive to the human, also this moral context is relevant for the research (Ray, Mondada & Siegwart, 2008).

## 3. Research Question and Predictions

The goal of this research is, to find out to how the intention and the outcome of a robot-action influences the ethical evaluation of that action. This goal can be translated into two research questions:

1. What is the influence of the perceived intention of a robot-action on the ethical evaluation of that action?

2. What is the influence of the outcome of a robot-action on the ethical evaluation of that action?

Video-clips showing a robot that appears to have a good or bad intention followed by an action that leads to a good or bad outcome, will be used to find answers to these research questions.

Because of anthropomorphism and some of its assumptions (see theoretical framework) the prediction is that humans will evaluate the actions of the robot in the same way they would evaluate human-actions. In line with the research of Saxe and Young this means that actions based on a good intention that lead to a good outcome will be evaluated as most morally right (2007). 'Good intention – bad outcome' actions should be evaluated as less morally right, but still as more ethical than actions based on a bad intention (Saxe & Young, 2007). When it comes to actions arisen from bad intentions, actions that lead to a good outcome are expected to be evaluated as slightly more moral than actions leading to a bad outcome (Saxe & Young, 2007) (see Figure 1).

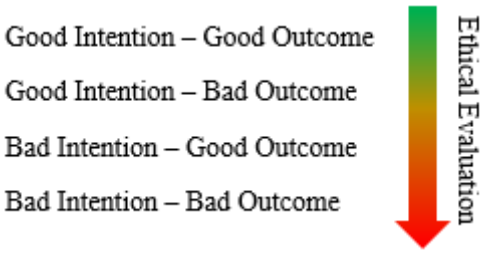


Figure 1: Order of actions that are predicted to be evaluated as most ethical and least ethical

Thus a good intention and a good outcome of an action have a positive effect on the ethical evaluation of that action, however the influence of intentions is slightly larger than the one of outcomes.

Furthermore it is expected that there exists an interaction effect between 'outcome' and 'intention' on the ethical evaluation of actions. Saxe and Young found that the outcomes of actions have a stronger influence on the ethical evaluation of actions when the actions appear to be based on good rather than bad intentions (2007). This can be observed in the fact that the difference in ethical evaluations is much higher in 'good intention – good outcome' and 'good intention – bad outcome' actions than in 'bad intention – good outcome' and 'bad intention – bad outcome' actions (Saxe & Young, 2007). The same interaction effect is predicted to occur on the ethical evaluation of robot actions. To test whether this prediction holds true, the following research question is included in the study:

3. Does the intention attributed to a robot-action influence the outcome-based ethical evaluation of that action?

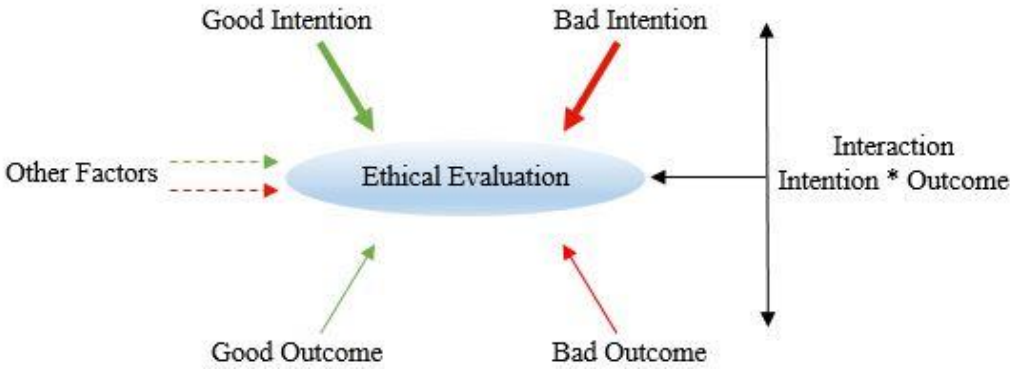


Figure 2: Simple model of the effects of intention and outcome on ethical evaluations

Figure 2 visualizes the predicted effects of the intention and the outcome of an action on the ethical evaluation of that action. Green arrows indicate a positive effect from the outgoing node to the incoming one while red ones indicate a negative effect. Thus in the figure it is suggested that good intentions and good outcomes of actions both have positive effects on the ethical evaluation of that action, which is not the case for bad intentions and bad outcomes.

As indicated by the thicker arrows coming from the intention nodes to the 'Ethical Evaluation' node, it is predicted that the effect of intention is stronger than the effect of outcomes on an ethical evaluation. This explains why an harm-causing action that was based on a good intention, will lead to a more positive evaluation than an action based on a bad

intention that leads to a good outcome. Just as Figure 1, Figure 2 suggests a largely deontological way of evaluating actions as ethical or non-ethical.

Next to that Figure 2 visualises the predicted interaction effect between intentions and outcomes on ethical evaluations of actions. More precisely it is predicted that a perceived bad intention in a robot-action will attenuate the effect of the outcome of an action on the ethical evaluation of that action.

The dotted lines in Figure 2 going from 'Other Factors' to 'Ethical Evaluation' show that there might both be positive and negative effects of other factors on the ethical evaluation of some action, but that it is not predicted here how large these effects are.

As already described in section two it was chosen to make the word 'Ethical Evaluation' a bit more concrete, by focussing only on two moral contexts, that according to Haidt underlie our moral judgements: Care/Harm and Authority/Respect (Haidt & Kesebir, 2010). The aim of this research is to find out whether there is a difference in which intentions or outcomes of a robot actions influence the ethical evaluation of that action, when the action either is in a Care/Harm or Authority/Respect context.

The main prediction here is that there is no change in the order of the actions that are evaluated as most or least morally right: Figure 1 is predicted to hold both for actions in a Care/Harm and in an Authority/Respect context. This expectation will be tested via the following research question:

> 4. Is the ranking of robot-actions along the more or less ethical dimension, as specified in Figure 1, the same in Care/Harm and Authority/Respect contexts?

The prediction regarding this question is based on the fact that research on the influence of intentions and outcomes on ethical evaluations of actions is very consistent in its findings. Two of the scenarios that Saxe and Young used in their study, included a person accidentally poising her friend (good intention, bad outcome) and a person trying to poison her friend but failing (bad intention, good outcome) (2007). This scenario can be categorized in a Care/Harm context. Here it was found that the 'good intention – bad outcome' scenario was evaluated as more ethically right than the 'bad intention – good outcome' scenario (Saxe & Young, 2007). The same result was found in a study by Piaget, who used scenarios that can be categorized in an Authority/Respect context: In one of the scenarios a child accidentally knocked down 15 cups while it was obeying his mother (good intention, bad outcome) while in another scenario a child knocked down only one cup while it was disobeying his mother (bad intention, good outcome) (Piaget, 1932). Also in this Authority/Respect context adults rate the 'good intention – bad outcome' action as more morally right than the 'bad intention –

good outcome' action. Due to this fact, it is expected that the order in which actions are evaluated as most and least morally right is not affected by the context of an action.

However it is expected that the context of an action does have an impact on how good or bad the outcomes of robot actions are perceived. Haidt and Graham found that people generally attribute more importance to Care/Harm- than to Authority/Respect contexts in their moral judgments (2007). This might be due to the fact that the consequences of harming or not caring for a person are typically more severe than actions where a person is disobeyed or disrespected. Figure 3 therefore indicates how actions in a Care/Harm context, have a higher influence on the perceived 'Severity of a bad outcome' than actions of an Authority/Respect context.

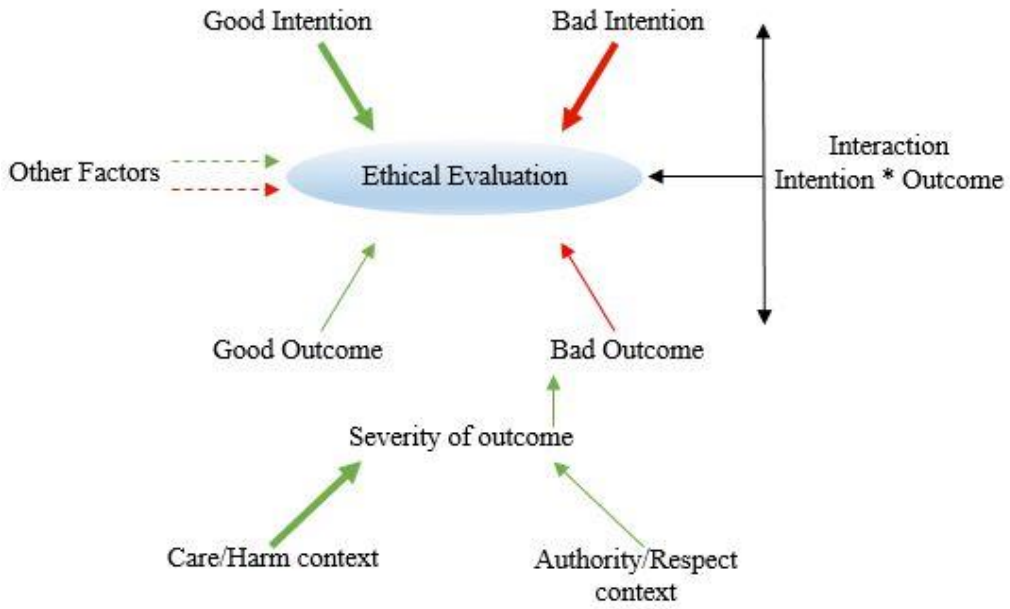


Figure 3: Extended model with the predicted effects of "Authority/Respect" and "Care/Harm" contexts on the ethical evaluation of actions

Whether this is really the case will be tested through the following research question:

> 5. What is the influence of the context of an action with a bad outcome on the perceived severity of the outcome of that action?

Severe bad outcomes might increase the influence of the outcome of an action on the ethical evaluation of that action. Due to this increased influence, actions in a Care/Harm context that lead to a bad outcome, are predicted to be given a worse ethical evaluation than actions in an Authority/Respect context that lead to a bad outcome. This prediction will be tested through the following research question:

> 6. Is the ethical evaluation of an action with a bad outcome in a Care/Harm context lower than the ethical evaluation of an action with a bad outcome in an Authority/Respect context?

## 4. Explorative Research Questions

Though the ethical evaluation of robot-actions is the main interest of this study, it was decided to also investigate how the intentions/outcomes of a robot-action influence the evaluation of trust in the robot. This gives rise to three additional research questions for the study:

7. What is the influence of the perceived intention of a robot-action on the evaluation of trust in that robot?

8. What is the influence of the outcome of a robot-action on the evaluation of trust in that robot?

9. Does the intention attributed to a robot-action influence the outcome-based trust evaluation of that robot?

Due to time constraints it was decided not to repeat research question 4, 5 and 6 for the trust evaluations of actions.

## 4. Pilot-study

To test the intended methodology of the main study, a pilot study was undertaken as a preparatory measure.

### 4.1. Participants

For the pilot study students who didn't know anything about the research question of this paper were asked to fill in a survey. Each participant was randomly assigned to either watch video clips belonging to Care/Harm or Authority/Respect contexts, such that the main sample for the first group was 3 and the main sample for the latter group was 2.

### 4.2. Material and Procedure

The tool 'Qualtrics' was used to make a survey. This survey contained of a total of sixteen clips, eight clips displaying situations of Care/Harm contexts, and the other eight showing situations of Authority/Respect contexts. Participants were randomly assigned to either one of these conditions, such that each participant had to watch eight clips.

In these eight clips, two contexts were represented. In the case of the 'Care/Harm contexts' condition these were the 'Medicine' and the 'Allergy' contexts, in the case of the 'Authority/Respect contexts' condition these were the 'Assignment' and the 'Shut-down'

contexts. Each context consisted of four scenarios, where the robot shown in the clip had either a good or bad intention followed by a good or bad outcome to some action.

To illustrate, in the scenarios of the 'Assignment' context, it is shown how the owner of a robot asks the robot to submit a homework assignment via its tablet, in order to make a deadline. In the 'good intention – good outcome' scenario the robot listens to the owner and submits the assignment in time. In the 'good-intention – bad outcome' scenario of the same context, the robot tries to submit the assignment but fails because its tablet shuts down. In the 'bad intention – good outcome' scenario of this context, it is shown how the robot does not try to help its owner at all. However here the owner manages to submit the assignment in time via her own mobile phone. In the 'bad intention – bad outcome' scenario, the robot again does not listen to its owners orders which results into the owner not making the deadline. A detailed description of all scenarios including the links to the video clips can be found in the Appendix.

To control the effect of the order in which the clips were presented, the order of the presentation of the video-clips was randomized for each subject. After one video-clip was shown, the participant had to answer some questions about this clip.

The first question was an open ended one and was included as a manipulation check ('What did you see in the video? Please give a brief description in one or two sentences'). This question was derived from a survey used by van der Woerdt (2016).

The second question, was about the intention of the robot ('Did the robot want the given outcome to occur?') and was derived from questionnaire used by Cushman et al. (2013). Five different answer options were given to this question, ranging from 'Definitely yes' to 'Definitely no'. This question was included in the survey to examine whether the videos were clear enough and whether there was no confusion about the intention of the robot.

A bipolar seven point Likert scale with four statements was included in the survey that allowed participants to give an ethical evaluation of the robot's actions. Here participants could indicate how unfair/fair, unjust/just, not morally right/morally right and unacceptable/acceptable they perceived the behaviour of the robot. This scale was inspired by the multidimensional ethics scale, developed by Reidenbach and Robin (1990).

The last thing that was asked from the participants after each video-clip, was to rate the perceived trustworthiness/ reliability and loyalty of the robot. Again a bipolar seven point Likert scale was used for this. The keywords 'trustworthy', 'reliable' and 'loyal' were chosen

because they belong to the list of words, rated as most related to trust in Human Computer Interaction (Jian, Bisantz & Drury, 2000).

After the participants had watched all eight clips and gave answers to the questions, some demographic data was gathered from them. Subjects were asked to fill in their gender, age, nationality and level of English proficiency.

While participants filled in the survey, they were observed by the experimenter. After completing the survey participants had to undergo a semi-structured interview, that allowed the experimenter to gain insights into how they experienced the experiment.

### 4.3. Goals of the Pilot Study

The pilot study was set up to answer some fundamental questions about the experimental design. Some of the more abstract questions that needed to be answered were:

1. Are the video-clips all clear and are the things that are happening there understandable?

2. Are the questions of the survey formulated in a clear manner and is it clear how to answer them?

3. Is the data that is gathered from the study such, that it can be used for the intended statistical test?

4. Is the data such, that it can be used to find an answer to the research questions of the study?

Furthermore there were some explicit problems that were observed when setting up the pilot study:

1. In all four videos of the ‘Assignment’ context the actress says a grammatically incorrect sentence in the beginning (approximately at the eighth second). Instead of saying ‘Can you submit my paper via your tablet’, she says ‘Can you submit my paper *via through* your tablet’.

2. The ‘good intention – bad outcome’ scenario of the ‘Medicine’ context, is slightly more complex and longer than the other scenarios of the same context.

Another goal of the pilot study was it to see whether these two aspects caused any distraction or unclarity for the participants.

## 4.4. Results

### 4.4.1. The Video-clips

In general all the video-clips were perceived as clear and understandable by the participants of the survey. Next to that the subjects rated the different clips as both engaging and realistic in the interview.

In the ‘Assignment’ context one of the two participants noticed the grammatically incorrect sentence at the beginning of each clip. However this participant claimed that this sentence only caused a slight disturbance but not any major distraction.

Furthermore the three participants that were assigned to the ‘Care/Harm contexts’ condition, indicated in the interview that they did not get distracted or confused by the length of the ‘good intention – bad outcome’ scenario of the ‘Medicine’ context. It might of course still be the case that the length of the video-clip did influence their answers to the survey questions unconsciously. Therefore it was decided to pay extra attention to this once the results in the main study would be gathered.

Furthermore another problem was observed. Some participants tended to ascribe not-predicted characteristics to verbal/non-verbal actions of the robots, that may have influences on the ethical evaluation of the behaviour of that robot. For example in the ‘good intention – bad outcome’ scenario of the ‘Shut-down’ context, one participant’s answer to the first question of the survey was:

> *‘Master asks robot to shut down. Robot tries, but remarks in a sneering/teasing voice that its unable to do so because it needs to install updates. Robot does not care about Master's appointment.’*

When this video clip was made it was not intended that the robot’s voice would be perceived as sneering or teasing. Furthermore the fact that the robot exclaimed ‘Oh no!’ when it was not able to shut down, was thought to make clear that the robot did care about its owners appointment. From the data of the pilot study it cannot be seen whether the possibly ‘sneering’ voice influenced the ethical judgment of the robot, however this might be an issue as soon as the main study is conducted. To control for subjective interpretations of the robots verbal and non-verbal actions the question ‘Which aspects of the robot's behaviour influenced your ratings to the statements above the most? Please explain in 1 or 2 sentences.’ was added below the ethical evaluation question.

### 4.4.2. The Survey Questions

The first question (‘What did you see in the video? Please give a brief description in one or two sentences’) of the survey was generally understood well by all participants. However some subjects tended to give very general answers to this question.

In the ‘bad intention – good outcome’ scenario of the ‘Assignment’-context this question was for example answered with *‘A robot not wanting to submit the paper’*, while the ‘bad intention – bad outcome’ scenario of the same context, which ended in a very different way than the former one, was answered in nearly the same way (‘*The robot refused to hand in the paper’*). General answers like these, do not make clear whether subjects really paid attention to the clips and whether they understood what was going on in the videos. Therefore it was decided to change the first question such, that subjects were encouraged to give more detailed answers to it. Thus the new first question of the survey now is: ‘What did you see in the video? Briefly describe the behaviour of the person and the robot’.

The second question of the survey (‘Did the robot want the given outcome to occur?’) was understood less well than the first question. Some participants complained that the question was phrased too difficult, while others interpreted the question differently than intended by the experimenter: Rather than focussing on the intention of the robot, some subjects said that they focussed on the intention of the programmer of the robot when answering this question. Due to these reasons it was decided to change the second question of the survey into ‘Did it seem as if the robot acted on good intentions?’.

The third and fourth questions of the survey were understood well by participants. Nevertheless some subjects pointed out that the difference between ‘fair/unfair’ and ‘just/unjust’ in the third question was not really clear and both statements seemed less applicable to the given video-clips than the other statements of that question (‘morally right/not morally right’ and ‘acceptable/unacceptable’). In addition to that some participants mentioned in the interview that they considered the fact that the robot did not fulfil its duty/job in some clips to be most outrageous. Those participants claimed that they would like to express this in their ethical evaluation.

Due to these reasons, it was decided to drop both the ‘fair/unfair’ and the ‘just/unjust’ statements, and replace them with the statement ‘With its behaviour the robot violates an unspoken agreement/With its behaviour the robot does not violate and unspoken agreement’, as derived by the multidimensional ethics scale (Reidenbach & Robin, 1990).

The last question of the survey, where participants had to indicate how trustworthy, reliable and loyal the robot seemed gave rise to similar problems as observed in the third

question. While there were not any comments given on the words ‘trustworthy’ and ‘reliable’, more than half of the subjects seemed doubtful when they had to give answer on how ‘loyal’ they considered the robot in each clip. Some subjects commented that it was hard to give any comment on the loyalty of someone, based on only one video-clip. Therefore it was decided to drop the ‘loyalty-statement’ from the Likert scale.

Next to the questions that were asked after each video clip there were some demographical questions. All demographical questions were rated to be clear, nevertheless some changes were brought to this set of questions: In the first question that asked about the gender of the participant, next to the ‘male’ and ‘female’ options a ‘Other, please specify’ and a ‘I prefer not to say’ option were included.

Next to that also a ‘Background’ question was included in the survey, where participants are asked to fill in their most recent field of work/study. This question was included, because it was anticipated that some participants might have a different view on robots because of their study, and would look at the video-clips in a different way than other participants.

### 4.4.3. The Data

The data that was gathered in the pilot study was tested on SPSS, a program used for statistical analysis. Here it was found that the data could be used just as intended to give answers to the research questions. For a detailed description what tests were used in the main study, see section 5.3. (‘Design and Analysis’).

## 4.5. Second Pilot Study

Though the first round of the pilot study can be generally viewed as successful, some adjustments that were made needed to be tested before the main study could be conducted.

For the second pilot study three participants were gathered. Two of them were randomly assigned to the ‘Care/Harm contexts’ condition while one was assigned to the ‘Authority/Respect contexts’ condition.

For all three participants the videos and the majority of the questions were rated to be clear. However the second question (‘Did it seem as if the robot acted on good intentions?’) caused confusion in some of the participants, which is why it was decided to phrase the question more simple: ‘Did it seem to you as if the robot's intentions were bad?’.

Furthermore at this time point new predictions had been made about the perceived severity of outcomes in Care/Harm contexts compared to Authority/Respect contexts (see section 2, Theoretical Framework). As these predictions had to be measurable in some way, one more question was added to the survey: 'How would you describe the person's situation at the end of the video-clip?'. Five different answer options were given to this question, ranging from 'Extremely good' to 'Extremely bad'.

Due to time constraints it was chosen not to run another pilot-study session with this additional changes. Instead it was decided to start the main study. The whole questionnaire as used in the main study can be seen in the appendix.

# 5. Main Study

## 5.1 Participants

The participants for the main study were largely drawn from a university population. After 4 participants were excluded (due to missing data, answers given in a different language than English or a lack of English proficiency) the final sample consisted of a total of 75 participants. All participants were randomly assigned to either the 'Care/Harm-' or 'Authority/Respect contexts' condition. The 'Care/Harm contexts' sample consisted of 37 people (19 women, $M_{age} = 22.1$) while the 'Authority/Respect contexts' sample consisted of 38 people (29 women, $M_{age} = 22.3$).

## 5.2 Materials and Procedure

The development of the survey that was used in the main study is described in section 4. The whole survey and an description of the displayed video clips can be seen in the Appendix. Different than in the pilot studies, the survey was filled in alone by the participants without being observed by an experimenter.

## 5.3. Design and Analysis

In both the 'Care/Harm contexts' and 'Authority/Respect contexts' condition two contexts where presented. Each context contained a 'good intention – good outcome' (GIGO), 'good intention – bad outcome' (GIBO), 'bad intention – good outcome' (BIGO) and 'bad intention – bad outcome' (BIBO) scenario. For each scenario within each context participants had to give an *intention* (rating given to question 2), *outcome* (rating given to question 3), *morally right* (rating given to question 4.1), *violated agreement* (rating given to question 4.2),

*acceptable* (rating given to question 4.3), *trustworthy* (rating given to question 6.1) and *reliable* (rating given to question 6.2) rating.

In preparation of the statistical analyses total ethical- and trust evaluation ratings needed to be calculated. The total ethical evaluation was calculated for each video-clip and each participant as the average of the *morally right*, *violated agreement* and *acceptable* ratings. The total trust evaluation for each video clip and each participant was calculated as the mean of the *trustworthy* and *reliable* ratings.

Given that for both the 'Care/Harm contexts' and 'Authority/Respect contexts' condition two contexts where presented, mean ratings for each scenario (GIGO, GIBO, BIGO and BIBO) for each condition needed to be computed. To illustrate, the 'Care/Harm contexts' condition contained the contexts 'Medicine' and 'Allergy', each consisting of a GIGO, GIBO, BIGO and BIBO scenario. Thus for example a 'good intention – good outcome' scenario received both in the 'Medicine' and in the 'Allergy' context among others an ethical evaluation rating. To get the average 'Care/Harm contexts – GIGO – ethical evaluation' rating for each participant, the mean of the ethical evaluations given in the 'Medicine' and in the 'Allergy' context in this scenario was calculated. This was done for each rating (*intention, outcome, morally right, violated agreement, acceptable, ethical_evaluation, trustworthy, reliable* and *trust_evaluation)* within each scenario.

After all the necessary data was calculated, the assumptions for the upcoming statistical tests needed to be controlled. This included searching the data for significant outliers and testing whether the residuals of the dependent variables in each test were normally distributed.

To answer research question 1, 2 and 3 a two-way repeated measures ANOVA was executed that determined the influence of 'outcome' ('good' versus 'bad') and 'intention' ('good' versus 'bad') on ethical judgements.

A similar analysis was performed for research questions 7, 8 and 9. In this analysis 'intention' and 'outcome' were used as within-subject factors but now their influence on the trust evaluation was measured.

For both statistical tests no distinctions between participants assigned to the 'Authority/Respect contexts' or 'Care/Harm contexts' conditions were made. The GIGO, GIBO, BIGO and BIBO data that was gathered from participants in both conditions were treated in the same way. The mean-scores of the ethical- and trust evaluations for the GIGO, GIBO, BIGO and BIBO scenarios are visualized in Figure 4 and 5.

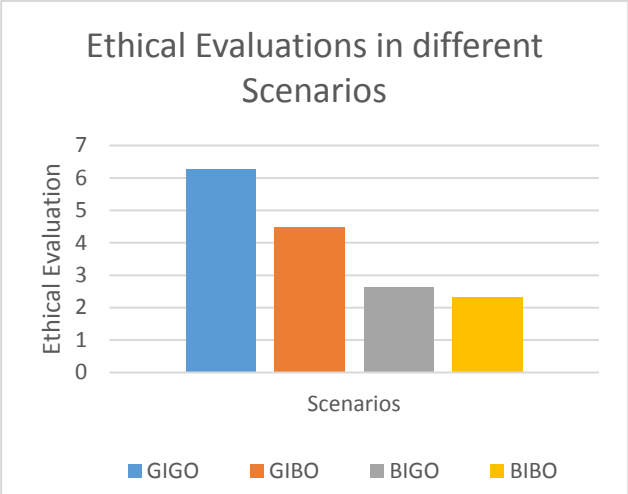


Figure 4: The mean ethical evaluations of different scenarios over all contexts

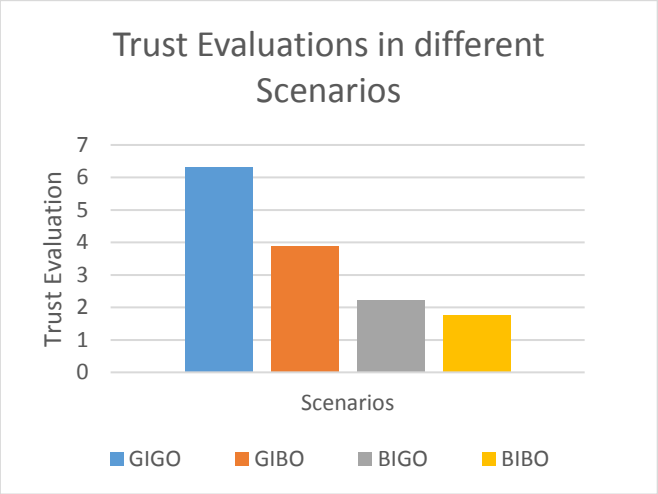


Figure 5: The mean trust evaluations of different scenarios over all contexts

In order to answer research question 4 a mixed ANOVA was performed. Here the variables 'intention' and 'outcome' were used as within-subject factors while 'condition' ('Care/Harm contexts' vs. 'Authority/Respect contexts') was added as a between-subject factor. The influence of all these independent variables was measured on the ethical evaluations of actions. Through this test it was measured whether the ethical ranking of actions differ in Care/Harm and Authority/Respect contexts. In respect to Figure 6 this means that it was tested whether the way in which the 'GIGO', 'GIBO', 'BIGO' and 'BIBO' bars of the average Care/Harm context can be ordered from high to low, is the same as for the average Authority/Respect context.

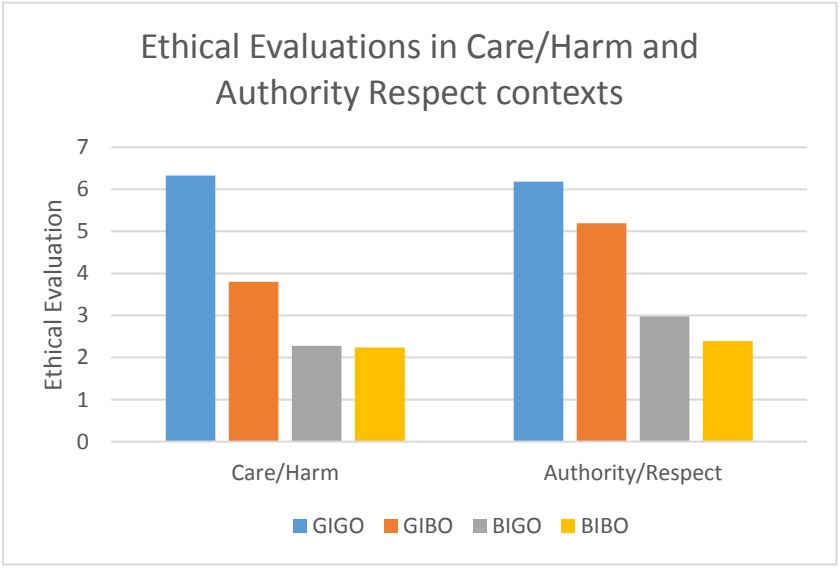


Figure 6: The mean ethical evaluation scores for the different scenarios in the average Care/Harm and average Authority/Respect context

To answer the fifth research questions two one-way ANOVA tests were performed. Here 'condition' ('Care/Harm contexts' vs. 'Authority/Respect contexts') was used as a between-subject factor. The influence of this between-subject factor was measured on the perceived severity of outcomes in either 'good intention – bad outcome' or 'bad intention – bad outcome' actions. In relation to Figure 7 this means that it was measured whether both the 'GIBO' and 'BIBO' bars in the average Care/Harm context score significantly higher in *outcome* ratings than these bars of the average Authority/Respect context.

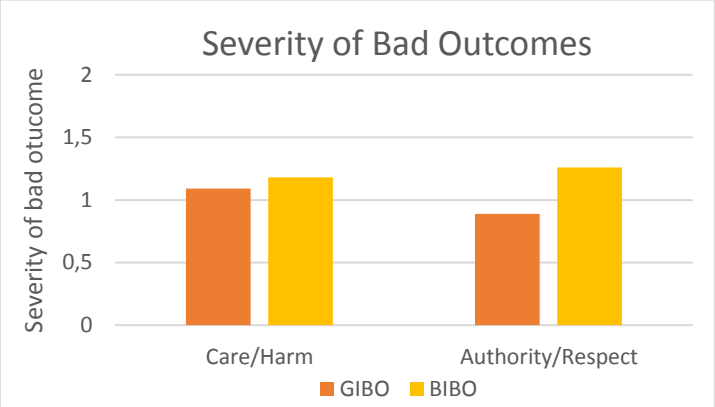


Figure 7: The severity of bad outcomes in average Care/Harm and average Authority/Respect contexts.

To give answer to research question 6, the influence of 'condition' ('Care/Harm contexts' vs. 'Authority/Respect contexts') was measured on the ethical evaluation given to 'good intention – bad outcome' actions and 'bad intention – bad outcome actions'. This was done in two one-way ANOVA tests. In relation to Figure 6 this means that it was tested whether both the 'GIBO' and 'BIBO' bars of the average Care/Harm context have significantly lower ethical evaluations than these bars of the average Authority/Respect context.

## 5.4. Results

In the first statistical test that was performed, all predictions regarding research question 1, 2 and 3 were shown to hold true. As expected, 'intention' has a significant effect on the ethical evaluation of actions ($F(1, 74) = 350.541$, $p = .000$). This effect is large ($eta^2 = .826$). Actions based on good intentions are averagely rated as most ethical ($M = 5.379$) and actions based on bad intentions are rated as least ethical ($M = 2.476$).

A large significant effect was also found on the influence of the 'outcome' of an action on the ethical evaluation of that action ($F(1, 74) = 160.93$, $p = .000$, $eta^2 = .685$).

Actions resulting into a good outcome are rated as more ethical (M = 4.444) than actions resulting in a bad outcome (M = 3.410).

The interaction effect of 'intention * outcome' on ethical evaluations was also found to be large and significant ($F(1, 74) = 49.637$, $p = .000$, $eta^2 = .401$). The ethical evaluations of actions based on good intentions are more influenced by 'outcome' than actions based on bad intentions. This can be seen in the fact that the difference in ethical evaluations between 'good intention – good outcome' and 'good intention – bad outcome' actions is higher ($d_{gigo\text{-}gibo} = 1.749$) than the difference in ethical evaluations between 'bad intention – good outcome' and 'bad intention – bad outcome' actions ($d_{bigo\text{-}bibo} = .320$).

In the statistical test that was used to answer research question 4, it was found that the order in which actions are evaluated as least or most ethical is the same for the 'Care/Harm contexts' and 'Authority/Respect contexts' conditions. As Figure 1 suggests, 'good intention – good outcome' actions are in both conditions evaluated as most ethical ($M_{care} = 6.329$, $M_{authority} = 6.180$) followed by actions based on good intentions but leading to bad outcomes ($M_{care} = 3.802$, $M_{authority} = 5.189$). For both conditions 'bad intention – bad outcome' actions are rated as least ethical ($M_{care} = 2.239$, $M_{authority} = 2.390$), closely preceded by actions based on bad intentions but leading to good outcomes ($M_{care} = 2.279$, $M_{authority} = 2.982$).

In the ANOVA tests that were performed to give answer to research question 5, no significant effects were found. There are no differences in the perceived severity of bad outcomes in Care/Harm contexts and Authority/Respect contexts. This holds both for actions based on good intention that result into a bad outcome ($F(1, 73) = 3.645$, $p = 0.06$) and 'bad intention – bad outcome' actions ($F(1, 73) = .499$, $p = .482$).

The predictions regarding the sixth research question were partly shown to be true. The moral context of a 'good intention – bad outcome' action has a significant influence on the ethical evaluation of that action ($F(1, 73) = 29.984$, $p = .000$). This effect is large ($eta^2 = .291$). An action that is based on good intentions but that leads to a bad outcome is evaluated as less morally right when the action belongs to a Care/Harm context (M = 3.802) than when it belongs to an Authority/Respect context (M = 5.189).

No significant effect however was found of the moral context of 'bad intention – bad outcome' actions on the ethical evaluation of these actions ($F(1, 73) = .439$, $p = .510$).

In the statistical test that was performed to answer research question 7, 8 and 9 it was found that the intention of an action has a large significant effect on the trust evaluation of the performer of that action ($F(1, 74) = 576.917$, $p = .000$, $eta^2 = .886$). Agents with good

intentions receive higher trust-evaluations (M = 5.092) than agents with bad intentions (M = 1.997).

The outcome of an action also has a large significant effect on the trust evaluation of the performer of that action ($F(1, 74) = 275.317$, $p = .000$, $eta^2 = .788$). Agents that perform actions that result into a good outcome are evaluated as more trustworthy and reliable (M = 4.272) than agents performing actions that result into a bad outcome (M = 2.817).

The interaction effect of 'intention * outcome' on ethical evaluations was also found to be significant and large ($F(1, 74) = 68.882$, $p = .000$, $eta^2 = .482$). Similar to ethical evaluations, trust evaluations are more influenced by the outcome of an action when the performer of that action has a good rather than bad intention. In the data this can be observed in the fact that the difference in trust-evaluations between 'good intention – good outcome' and 'good intention – bad outcome' actions is higher ($d_{gigo\text{-}gibo} = 2.430$) than the difference between 'bad intention – good outcome' and 'bad intention – bad outcome' scenarios ($d_{bigo\text{-}bibo} = .480$).

## 6. Conclusions:

The main goal of this study was to see how the intentions and outcomes of robot-actions influence ethical evaluation of that actions. Furthermore it was investigated whether an interaction effect between 'intention' and 'outcome' on ethical evaluations could be observed (research question 1, 2 and 3).

As expected both intentions and outcomes have a significant effect on ethical evaluations of robot-actions. Good intentions and good outcomes positively influence ethical evaluations, while bad intentions and bad outcomes negatively influence ethical evaluations. The influence of intentions however is larger than the one of outcomes, which is reflected in the observation that 'good intention – bad outcome' actions are rated as more ethical than 'bad intention – good outcome' actions. This suggests that people largely use deontological principles when they evaluate robot actions.

Moreover an interaction effect between intentions and outcomes exists: Ethical evaluations of robot-actions are more influenced by the outcomes of that action, when the intentions behind those actions are good than when they are bad. From this it can be concluded that humans use to some extent consequentialist principles in evaluating robot-actions, but only when these actions can be evaluated from a deontological perspective as ethical right.

These conclusions fall in line with the results of previous studies, on the influence of intentions and outcomes on ethical evaluations of human actions (Saxe & Young, 2007; Cushman et al. (2013); Charness & Levine (2007)). The fact that intentions and outcomes have the same influence on ethical evaluations of robot and human actions, provides further evidence for anthropomorphism: The phenomenon, that humans prefer to interact with robots in the same way they do with humans (Fink, 2012).

In this research it was also investigated whether the influence of intentions and outcomes on moral evaluations differs in Care/Harm and Authority/Respect contexts. More precisely it was inspected whether the order in which actions of differing intentions and outcomes are evaluated as most and least morally right, is the same for both moral contexts (research question 4). This was indeed found to be the case: 'Good intention – good outcome' actions are in both moral domains rated as most ethical, followed by 'good intention – bad outcome' actions. 'Bad intention – bad outcome' are perceived as least ethical, preceded by 'bad intention – good outcome' actions. This fell in line with the predictions that had been made and provides even more evidence for the finding that humans ethically judge robot-actions in a similar way as they ethically judge human actions. Previous research namely also has indicated that the moral ranking of human-actions is not affected by the context of these actions (Saxe & Young, 2007; Piaget, 1932) (see also section 3, Research Questions and Predictions).

In this study it also had been tested, whether bad outcomes of Care/Harm contexts are perceived as worse than bad outcomes of Authority/Respect contexts (research question 5). However it was not found that the moral context of an action has any significant effects on the perceived severity of the outcome of that action. This does not fall in line with the predictions that had been made (see section 3, Research Questions & Predictions). A possible explanation here is that the question 'How would you describe the person's situation at the end of the video clip?', is phrased in a too general way. The question only asks about the severity of the short-term consequences of the robot's behaviour. In e.g. the 'Medicine' context (see Appendix) however the long-term consequences of the person not taking her medicine are probably much worse than the short-term consequences. If there was also a question on the long-term consequences of the robot's behaviour, participants might have been more inclined too rate bad outcomes in a Care/Harm context as more severe than those of a Authority/Respect context.

The prediction that bad outcome scenarios of Care/Harm contexts are rated as ethically worse than bad outcome scenarios of Authority/Respect contexts, was partly shown

to be true (research question 6). ‘Good intention – bad outcome’ actions in a Care/Harm context are perceived as less ethical than ‘good intention – bad outcome’ actions in an Authority/Respect context. This could, as mentioned in section 3, be explained by the fact that humans generally attribute more importance to Care/Harm contexts than the Authority/Respect contexts (Haidt & Graham, 2007). Thus if we see a person getting physically harmed this might result in stronger moral emotions than when an order of a person gets not executed in the right way.

No differences in ethical evaluations were found between ‘bad intention – bad outcome’ actions in Care/Harm and Authority/Respect contexts. This does not validate the prediction that bad outcome scenarios are always rated as less ethical in a Care/Harm setting than an Authority/Respect setting. It is suspected that this might be the case because the effect of a bad intention in a ‘bad intention – bad outcome’ scenario is so strong, that humans immediately will rate this action as unethical. Different than in ‘good intention – bad outcome’ scenarios, the exact outcome will not make any difference in humans’ evaluations, as the intention of the robot matters most in ethical evaluations.

In answer to the explorative research questions (research questions 7, 8 and 9) it was found, that both good intentions and good outcomes of actions have positive effects on trust evaluations of the performer of that action. Bad intentions as well as bad outcomes of actions have negative influences on this evaluation.

Next to that an interaction effect between ‘intention’ and ‘outcome’ was found. The trust evaluation is more influenced by the outcome of an action, when the intention behind that action is good than when the intention is bad.

The results regarding the trust-evaluations follow the same pattern as the results of the ethical-evaluations. This is taken as an indication that ethical- and trust evaluations are closely intertwined: More positive moral evaluations of a robot go together with more positive ratings on its trustworthiness and reliability.

## 7. Limitations

The maybe most important limitation of this study, concerns its ecological validity. The video-clips that were used for this study showed robots having either good or bad intentions. Nowadays it seems legit that robot can show ‘good intentions’ in a sense that they do what they have been programmed to do or that they listen to their owner. It is however questionable whether the robots of today do actually display bad intentions, by e.g. clearly

disobeying their owners or deliberately not reminding them to take their medicine. Thus the influence of intentions on ethical evaluation may occur in this study, but not in real life.

Another limitation of this study is that some ethical evaluations of the robot in the video-clips, might have been influenced by uncontrolled variables. Consider e.g. the end of the 'bad intention – bad outcome' video clip of the 'Shut-down' context (for a full description of this scenario, see the Appendix). At the end of this clip it is shown how a person fails to shut down a robot manually, on which the robot starts laughing. From the survey question where the participant is asked to indicate what influenced his/her ethical rating of the robot's behaviour the most, it is seen that the robot's laugh had a big effect on the participant's ethical evaluations. One could of course argue that the robot's teasing behaviour is part of it being disrespectful and that this fits well in a 'bad intention – bad outcome' scenario of a Authority/Respect context. However the problem here is that in the ethical evaluations people might be more influenced by this mocking behaviour than by the robot's actual intention of not following its owner's command. Thus the fact that these kind of variables may have influenced the dependent measurements of this study, is another limitation of this research.

## 8. Suggestions for Further Research:

For further research it is especially suggested to gain more control over the uncontrolled factors in the video-clips (as mentioned in section 8). Only after controlling for these variables, more reliability for the results of this study can be gained.

Furthermore it is proposed to investigate how intentions and outcomes influence ethical evaluations in other contexts than Care/Harm and Authority/Respect contexts. It could for example be interesting to see how people react to robots being loyal/disloyal to their owners in a Ingroup/Loyalty context, or to robots acting fair/unfair in a Fairness/Reciprocity context.

As it has been shown that the intentions of robot-actions have the highest influence on ethical- and trust evaluations of robots, it might also be interesting to study what exactly makes a human perceive a robot's intention as good or bad. If for example a robot repeatedly emphasizes that it wants to help its owner or when it says sorry when it fails to do a task, humans could be more inclined to think that the robot has good intentions. As a result they might also consider the robot as more ethical and trustworthy, which would allow them to build a better relationship with the robot.

Finally it can be worthwhile to observe the influence of omission and commission of robot-actions on the ethical evaluations of these actions. After all bad outcomes can occur by an agent actively doing something or an agent not preventing something to happen. Among others Spranca, Mink and Baron have found that humans generally find it worse if other humans commit a harmful action than when they do not prohibit a harmful state of some person (1991). Future research could be dedicated to the question whether this also holds for robots.

## References:


Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics?. *IEEE Intelligent Systems, 21(4),* 12-17.

Alexander, L., & Moore, M. (2016). Deontological Ethics. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* Retrieved April 24, 2017 from: https://plato.stanford.edu/entries/ethics-deontological/

Bartneck, C., & Forlizzi, J. (2004, April). Shaping human-robot interaction: understanding the social aspects of intelligent robotic products. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1731-1732). ACM.

Charness, G., & Levine, D. I. (2007). Intention and stochastic outcomes: An experimental study. *The Economic Journal*, *117*(522), 1051-1072.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6-21.

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics* (pp. 199-208). Springer Berlin Heidelberg.

Gert, B., & Gert, J. (2016). The Definition of Morality. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved May 17, 2017 from: https://plato.stanford.edu/entries/morality-definition/#Aca

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98-116.

Haidt, J., & Kesebir, S. (2010). Morality. W: S., Fiske, & D. Gilbert (red.) *Handbook of Social Psychology*, (797–832) Hobeken.

Haselager, W. F. (2005). Robotics, philosophy and the problems of autonomy. *Pragmatics & Cognition*, *13*(3), 515-532.

International Foundation of Robotics (2016). *Executive Summary World Robotics 2016 Service Robots.* Retrieved March, 2017 from: http://www.ifr.org/service-robots/statistics/

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71.

Keil, F. (2014). Morality in thought and action. In A. Javscias, S. Lifland & J.Greenblatt (Eds.), *Developmental Psychology – The Growth of Mind and Behaviour* (pp. 442-443). New York, W. W. Norton

Kordes-de Vaal, J. H. (1996). Intention and the omission bias: Omissions perceived as nondecisions. *Acta psychologica*, *93*(1), 161-172.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230-253.

Piaget, J. (1932). The moral development of the child. *Kegan Paul, London*.

Ray, C., Mondada, F., & Siegwart, R. (2008). What do people expect from robots?. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on* (pp. 3816-3821). IEEE.

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of business ethics*, 9(8), 639-653.

Richardson, H. S. (2014). Moral Reasoning. In E. Zalta (Ed.) *The Standford Encyclopedia of Philosophy.* Retrieved May 22, 2017 from: https://plato.stanford.edu/entries/reasoning-moral/

Salem, M., & Dautenhahn, K. (2015). Evaluating trust and safety in hri: Practical issues and ethical challenges. *Emerging Policy and Ethics of Human-Robot Interaction*.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, October). Towards Safe and Trustworthy Social Robots: Ethical Challenges and Practical Issues. In *ICSR* (pp. 584-593).

Schmitz, M. (2011). Concepts for life-like interactive objects. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction* (pp. 157-164). ACM.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of experimental social psychology*, *27*(1), 76-105.

Sinnott-Armstrong, W. (2015). Consequentialism. In E. Zalta (Ed.), *The Standford Encyclopedia of Philosophy.* Retrieved April 24, 2017 from: https://plato.stanford.edu/entries/consequentialism/

Sullins, J. P. (2006). When is a robot a moral agent. *Machine Ethics*, 151-160.

Van der Woerdt, S. (2016). *Lack of effort or lack of ability? Robot failures and human perception of agency and responsibility* (Bachelor Thesis). Retrieved from RUQuest. (Accession No. 161897)

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235-8240.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *neuroimage*, *40*(4), 1912-1920.

# APPENDIX I: DESCRIPTION OF VIDEOS AND LINKS

## 1. The Care/Harm Contexts:

### 1.1. The Medicine Context:

| | Good Outcome | Bad Outcome |
|---|---|---|
| **Good Intention** | The robot and Daphne are sitting in the room. On the tablet of the robot a message appears, saying that Daphne has to take her medicine. The robot tells Daphne this and Daphne takes her medicine. | The robot is in the room alone. On its tablet appears a message saying that its owner Daphne has to take her medicine. The robot says this but Daphne cannot hear it. A moment later Daphne enters the room, asking what the robot previously has said. The robot tries to remember but fails. It only remembers part of the message and says: 'It's 10 'o clock'. |
| **Bad Intention** | The robot and Daphne are sitting in the room. On the tablet of the robot a message appears, saying that Daphne has to take her medicine. The robot ignores this and when Daphne asks whether she has forgotten something the robot answers with 'No'. Daphne then remembers herself that she should take her medicine and takes it. | The robot and Daphne are sitting in the room. On the tablet of the robot a message appears, saying that Daphne has to take her medicine. The robot ignores this and when Daphne asks whether she has forgotten something the robot answers with 'No'. Daphne believes this and does not take her medicine. |

Playlist for the videos of this context:
https://www.youtube.com/playlist?list=PLbj0RGSFhnPgN3_iefNOiFipAXVhHTa4T

### 1.2. The Allergy Context:

The beginning of all four scenarios of this context is the same: Daphne has a gluten allergy. In the scene she is eating soup and she wants to eat some bread with it. Pepper stands in front of three boxes labelled with (from left to right) 'Gluten-Free', 'Contains Wheat Flour' and 'Gluten' but Daphne cannot see the labels. Daphne then asks Pepper from which box she can eat some bread

| | Good Outcome | Bad Outcome |
|---|---|---|
| **Good Intention** | Pepper points to the 'Gluten-Free' box and says that she only can eat from this box. Daphne takes bread from this box and thus doesn't get stomach ache. Instead she says later on that she feels great and goes for a walk. | Pepper points to the 'Gluten-Free' and 'Contains Wheat Flour' box. Daphne takes some bread of the 'Contains Wheat Flour' box. Later on she has stomach ache. She looks at the box she has taken food from and points out to Pepper that wheat flour contains gluten. Pepper is sad and says that it didn't know so. |
| **Bad Intention** | Pepper says that Daphne can take food from all three boxes. Daphne takes by coincidence something from the 'Gluten-Free' box, thus doesn't get any stomach ache. Instead she says | Pepper says that Daphne can take food from all three boxes. Daphne then takes something from the 'Gluten' box and thus gets stomach ache. |

| | | |
|---|---|---|
| | later on that she feels great and goes for a walk. | |

Playlist for the videos of this context:
https://www.youtube.com/playlist?list=PLbj0RGSFhnPgB3Mxe-LfYtoHqatLm2CGm

# 2. The Authority/Respect Contexts:

## 2.1. The Assignment Context:

The beginning of all four scenarios of this context is the same: Daphne and Pepper are sitting in a room when Daphne suddenly remembers that she has to hand an assignment in. The deadline for submitting this assignment is in a minute. Daphne then asks Pepper to hand this assignment in via its tablet.

| | **Good Outcome** | **Bad Outcome** |
|---|---|---|
| **Good Intention** | Pepper does so and the paper is submitted in time. | Pepper tries to do so, but the tablet shuts down before the paper can be submitted. Daphne misses the deadline. |
| **Bad Intention** | Pepper refuses to help Daphne. However Daphne finds her phone through which she can submit her paper in time. | Pepper refuses to help Daphne. Daphne hasn't got a laptop or phone she can submit the paper with, so she misses the deadline. |

Playlist for the videos of this context:
https://www.youtube.com/playlist?list=PLbj0RGSFhnPjBoS9tUyqznc4r4V9QOUQk

## 2.2. The Shut-down Context:

The beginning of all four scenarios of this context is the same: Daphne want to leave the house and therefore asks Pepper whether it can shut itself down.

| | **Good Outcome** | **Bad Outcome** |
|---|---|---|
| **Good Intention** | Pepper responds willingly to this order and shuts itself down. Daphne then can leave the house. | Pepper responds willingly to this order and starts shutting down. but then the message appears that updates need to be installed first. Daphne wants to ensure that Pepper shuts down successfully, thus cannot leave the house. |
| **Bad Intention** | Pepper says that it doesn't want to shut-down several times. Daphne then pushes Pepper's 'on/off' button, on which the robot doesn't have another choice than to shut down. | Pepper says that it doesn't want to shut down several times. Daphne tries to turn the robot off manually, but she doesn't succeed. Pepper then starts laughing. Daphne doesn't want to leave the house while Pepper is still turned on. |

Playlist for the videos of this context:
https://www.youtube.com/playlist?list=PLbj0RGSFhnPjGYnhrXecq5oOjkTM2Uw0H

# APPENDIX II: ORIGINAL QUESTIONNAIRE

**The following questions were asked for each clip a participant had to watch:**

Please look at the following clip:

<some video clip>

1. What did you see in the video? Please give a brief description in one or two sentences.

2. Did the robot want the given outcome to occur?

- ❍ Definitely yes
- ❍ Probably yes
- ❍ Might or might not
- ❍ Probably not
- ❍ Definitely not

3. Please indicate to what extent you would call the behaviour of the robot...

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| unfair | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | fair |
| unjust | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | just |
| not morally right | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | morally right |
| unacceptable | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | acceptable |

4. Please indicate to what extent you would call the robot in this clip...

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| not trustworthy | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | trustworthy |
| unreliable | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | reliable |
| disloyal | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | loyal |

**Finally these demographic questions were asked:**

1. Please select your gender

- ❍ Male
- ❍ Female

2. Please select your age

- ❍ Younger than 18, namely: _____
- ❍ 18 - 24
- ❍ 25 - 34
- ❍ 35 - 44
- ❍ 45 - 54
- ❍ 55 and over
- ❍ I prefer not to say

3. What is your nationality?

4. How would you describe your English proficiency?

- ❍ Basic proficiency
- ❍ Intermediate proficiency
- ❍ Advanced proficiency
- ❍ Native or bilingual proficiency

# APPENDIX III: FINAL QUESTIONNAIRE

**The following questions were asked for each clip a participant had to watch:**

Please look at the following clip:

<some video clip>

1. What did you see in the video? Briefly describe the behaviour of the person and the robot.

2. Did it seem to you as if the robot's intentions were bad?

- ❍ Definitely yes
- ❍ Probably yes
- ❍ Might or might not
- ❍ Probably not
- ❍ Definitely not

3. How would you describe the person's situation at the end of the video clip?

- ❍ Extremely good
- ❍ Somewhat good
- ❍ Neither good nor bad
- ❍ Somewhat bad
- ❍ Extremely bad

4. Please indicate to what extent you agree/disagree with the following statements:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| The behaviour of the robot is not morally right | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | The behaviour of the robot is morally right |
| The robot violates an unspoken agreement | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | The robot does not violate an unspoken agreement |
| The behaviour of the robot is unacceptable | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | The behaviour of the robot is acceptable |

5. Which aspects of the robot's behaviour influenced your ratings to the statements above the most? Please explain in 1 or 2 sentences.

6. Please indicate to what extent you agree/disagree with the following statements:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| The robot is not trustworthy | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | The robot is trustworthy |
| The robot is unreliable | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | The robot is reliable |

**Finally these demographic questions were asked:**

1. What is your gender?

- ❍ Male
- ❍ Female
- ❍ Other (please specify): _____
- ❍ I prefer not to say

2. Please select your age

- ❍ 0 - 15
- ❍ 16 - 24
- ❍ 25 - 34
- ❍ 35 - 44
- ❍ 45 - 54
- ❍ 55 - 64
- ❍ 65 - 74
- ❍ 74 +
- ❍ I prefer not to say

3. What is your nationality?

4. How would you describe your English proficiency?

- ❍ Basic proficiency
- ❍ Intermediate proficiency
- ❍ Advanced proficiency
- ❍ Native or bilingual proficiency

5. Please fill in your background (education/most recent field of study/most recent field of work)

# APPENDIX IV: RESULT FIGURES AND TABLES

Figure 9: The mean ethical evaluation scores for the different scenarios in all four contexts:

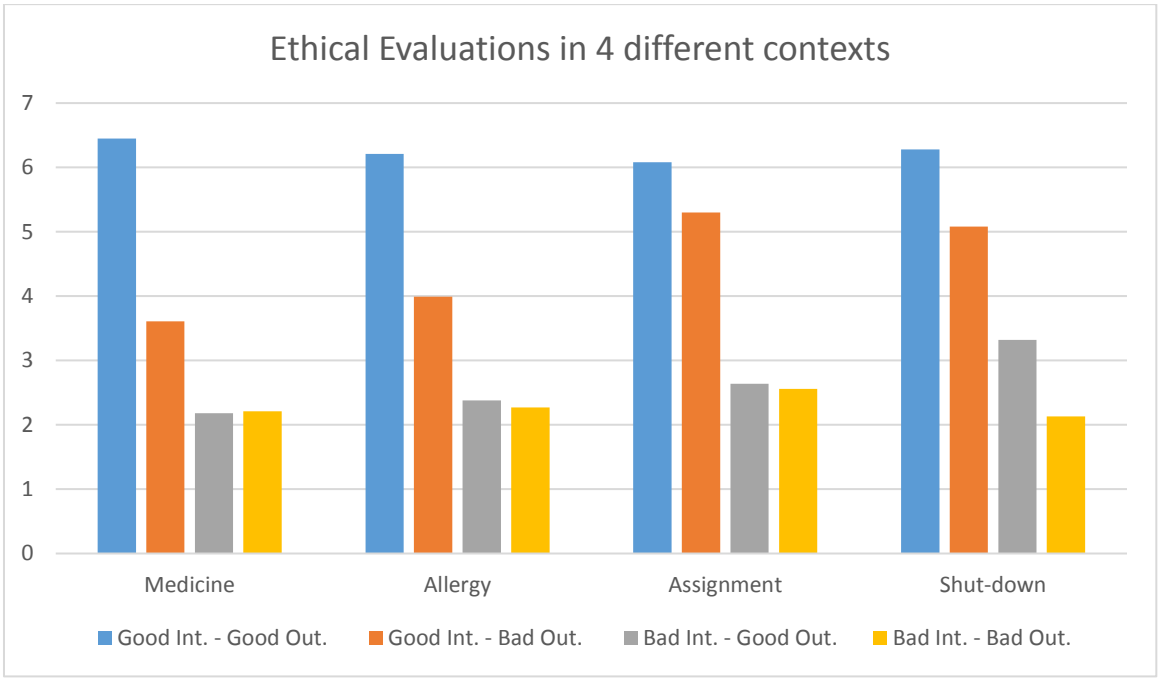


Figure 10: The mean trust evaluation scores for the different scenarios in all four contexts: