Lack of Effort or Lack of Ability? Robot Failures and Human Perception of Agency and Responsibility

Sophie van der Woerdt[1] and Pim Haselager[2](✉)

[1] Department of Psychology, Donders Institute for Brain,
Cognition and Behaviour, Radboud University,
Comeniuslaan 4, 6525 HP Nijmegen, Netherlands
[2] Department of Artificial Intelligence, Donders Institute for Brain,
Cognition and Behaviour, Radboud University,
Comeniuslaan 4, 6525 HP Nijmegen, Netherlands
w.haselager@donders.ru.nl

Abstract. Research on human interaction has shown that considering an agent's actions related to either effort or ability can have important consequences for attributions of responsibility. In this study, these findings have been applied in a HRI context, investigating how participants' interpretation of a robot failure in terms of effort -as opposed to ability- may be operationalized and how this influences the human perception of the robot having agency over and responsibility for its actions. Results indicate that a robot displaying lack of effort significantly increases human attributions of agency and –to some extent- moral responsibility to the robot. Moreover, we found that a robot's display of lack of effort does not lead to the level of affective and behavioral reactions of participants normally found in reactions to other human agents.

Keywords: Agency · Responsibility · Human-robot interaction · HRI · Attribution · Anthropomorphism · Mind perception · Social cognition · Theory of Social Conduct

1 Introduction

Even when built to be perfect, computer- and robotic-systems are known to occasionally malfunction in or throughout the task they were designed to perform. Although in practice the consequences of malfunction, misuse or mere accidents with the usage of robots are oftentimes innocent (e.g. a household robot dropping a plate on the ground), one can also think of consequences that cause more serious harm (e.g. a

This paper is based on a thesis that was submitted in fulfilment of the requirements for the degree of Bachelor of Science (Honours) in Psychology at the Radboud University in August 2016. Another paper based on this thesis has been submitted to the journal *New Ideas in Psychology* [1], focusing primarily on the use of theories in social psychology for HRI. The current paper focuses specifically on the implications of Weiner's theory for the implementation or robot behavior aimed at eliciting different attributions of agency and responsibility.

T. Bosse and B. Bredeweg (Eds.): BNAIC 2016, CCIS 765, pp. 155–168, 2017.
DOI: 10.1007/978-3-319-67468-1_11

household robot dropping a plate on a pet or a child). In fact, considering the increasing number of applications of robots, and the increasing number of people using them, it will be impossible to predict all the different ways in which robots will be used and the mistakes that robots can make.

Hence, currently, much debate is devoted to the question of how we should deal with such harms caused by robots [2, 3]. One central issue in this discussion is the role of anthropomorphism, that is: the human tendency to assign human traits, emotions and intentions to non-humans. More specifically, if we assume robots do not think or feel, it would be impossible to blame them for their acts, let alone to punish them. Nevertheless, while in theory people remain that robots are no candidates for inferring thoughts and feelings to, in practice research on anthropomorphism shows that humans automatically take perspective when seeing the movements of inanimate objects [4–7]. This can frequently be noticed in daily life, for example in the way we are emotionally involved in watching animated films, socially connect with stuffed animals, or speak of weather or a stock market that is ‘pleased’ or ‘angry’ [8].

Considering our tendency to think about and/or act towards robots as if it were humans, we believe mechanisms of human-human interaction may also be applicable in human-robot interaction (HRI). One such mechanism has been comprehensively studied in Weiner’s *Theory of Social Conduct*, describing the precursors and consequences of humans attributing agency and responsibility to other agents. Weiner’s studies [9] reveal that especially attributions of controllability and uncontrollability matter in perceiving whether people act with *agency* (having an autonomous or at least partially independent capacity to engage in goal-directed action) or not, and consequently whether these people bear *responsibility* (whether one can be praised or blamed for its actions). This in turn is shown to have effect on fundamental emotional and behavioral responses such as acceptance, rejection, altruism and aggression. For example, in one of their studies, Weiner et al. [10] let participants judge hypothetical patients suffering from diseases that are generally considered as more controllable than others (e.g. obesity and drug abuse are perceived as controllable; cancer and Alzheimer as uncontrollable), and found that patients suffering from ‘controllable diseases’ are judged as carrying more responsibility for their condition. In addition, results indicated that the people suffering from these ‘controllable diseases’ were less likely to be helped, receive donations, or even just being liked as a person. These findings have been replicated in and with regard to different contexts such as school settings, stigmas, and reactions to penalties for an offense [9, 11, 12]. Therefore, given robots’ potential to cause undesired outcomes, combined with the tendency of human users to anthropomorphize them, we suggest that robots may also be judged under the influence of attributions of controllability.

In addition to the general likeability and acceptance of robots in daily life situations, we believe there could be even more at stake. Anthropomorphism may cause owners and developers to (unknowingly) distance themselves from potential harms caused by their robots [13], causing responsibility to become diffused. Thus, we believe that finding out more about attributions of agency and responsibility in robots is of great societal relevance.

Nevertheless, with regard to HRI, little is known about attributions of agency and moral responsibility. In fact, the topic of responsibility in HCI/HRI has been incorporated in a number of studies, but these do not necessarily reflect attributions of agency or moral responsibility via attributions of a mind to the computers or robots involved (for a more extensive report on this, see [1]). For example, in the context of a collaborative game setting, Vilaza et al. [14] found that participants have a tendency to blame computers or robots when a game is lost. Similar results were reported by Moon and Nass [15] and You et al. [16], who found that participants tend to blame computers when their results were evaluated negatively while taking credit when a game is won or when results are evaluated positively. Hence, in these cases, robots are primarily blamed as a consequence of self-serving bias rather than as a consequence of anthropomorphism. Moreover, in two recent studies Malle et al. [17, 18] presented participants with a picture of either a mechanical robot or a humanoid robot responding to moral dilemmas. Their results showed that, as compared to the mechanical robot with regards to judgments of blame, participants judged the humanoid robot more similar to how a human agent would be judged. Nevertheless, there still remains a lot of unanswered questions as to what incites attributions of agency and responsibility in HRI, and how exactly this could be operationalized for the benefit of better HRI.

An important factor that influences the distinction between controllable and uncontrollable outcomes is the perception of the amount of *ability* and *effort* that is displayed in behavior. For example, in school settings, despite the grade-outcome, teachers prefer to praise their students in terms of how much effort a student puts into the work. So students with low ability (uncontrollable cause), but high motivation (controllable cause) are often considered as better or at least more likeable students than students with high ability but low motivation [19]. Considering malfunctions in robotic behavior, we think these attributions of ability and effort may be well applicable in HRI. Therefore, in this study we performed a small experiment in which participants were shown videos of robots (Aldebaran's NAO; https://www.ald.softbankrobotics.com/en) failing tasks in ways that could be interpreted as due to either *lack of ability* (LA-condition; e.g. dropping an object) or *lack of effort* (LE-condition; e.g. throwing away an object).

Accordingly, the main dependent variables in our study were defined as *agency* and *responsibility.* For our purposes, we define agency as the attribution of the capacity to act towards the realization of a goal, and responsibility as being accepted as a candidate for the attribution of credit or blame. It is important to note that attributing agency does not necessarily imply the attribution of responsibility [20, 21], for example in the case of children, subordinates following orders during their job, people with (temporary) diminished mental capacity or even domesticated animals. In all these occurrences, agents display agency, but do not carry full responsibility due to the presence of mitigating circumstances (e.g. not knowing right from wrong or inability to comport behavior to the requirements of law; [2]). Some authors have drawn parallels between robots and domestic animals in this regard, recognizing that both are often afforded similar capacities, rights and responsibilities [22, 23].

Agency and responsibility also need to be distinguished from another feature that can be attributed to robots, namely: *experience* (e.g. the attribution of an agent having beliefs, intentions, desires, emotions). Over the years, in most research anthropomorphism has been loosely defined as "the assignment of human traits, emotions and intentions to non-humans". However, a factor analysis of Gray et al. ([24], but also see [25–31]) revealed that in this regard we can better speak of two individual factors of mind perception: experience and agency. Therefore, in order to get a better idea of what type of mind perception we are actually measuring, we included measuring attributions of experience in our data collection. However, it should be noted that this distinction is not the main focus of our study.

In fact, the main goal of our experiment was to examine the effects of showing videos of robots failing due to lack of ability or due to lack of effort on attributions of agency and responsibility. We expected that a display of lack of effort would incite the illusion of a robot having agency over its actions. Contrarily, we did not formulate definite predictions about the possible effects of our robots displaying lack of effort on measures of responsibility. Since these conditions -especially the LE-condition- were supposed to represent situations that in fact depend on subjective evaluations, in this paper we spent special attention to the operationalization of the LA/LE-conditions. Therefore, as a secondary goal, we considered the extent to which each of the videos separately affects attributions of agency, responsibility and experience.

Devising robot behavior to elicit attributions of agency, responsibility and experience was not straightforward. Several constraints had to be met. First, it should be clear to participants what task the robot is trying (or not trying) to do, even without seeing successful task-performance. Second, it should be clear to participants that failure of task-performance is indeed a failure, that is: it should at least have some negative consequences. On the other hand, assuming robots in real life will not be programmed to do morally objectionable behavior, the robots' behavior should at most passively cause harm (i.e. negligence, not 'trying'), but not actively (i.e. hurting, cheating, lying for one's own benefit). Third, since we wanted to limit other influences than mere task performance as much as possible, we tried to limit narrative and dialogue. Fourth and finally, both for practical reasons as well as for the generalizability of the study: the tasks should be simple enough for a NAO robot to perform.

2 Method

Participants. Participants were drawn from a university population in exchange for a €5 gift-certificate. After listwise exclusion of eight participants (due to missing data and double responses) the final sample consisted of a total of 63 participants. These participants were randomly divided amongst the LA- and LE-conditions. The LA-sample consisted of 31 people (19 women, M_{Age} = 26,7, SD = 11,6). The LE-sample consisted of 32 people (14 women, M_{Age} = 26,3, SD = 7,5).

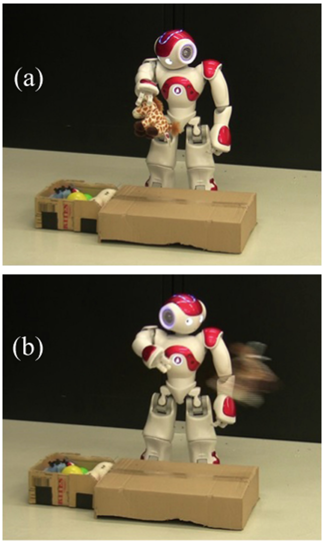

Fig. 1. Sample frames of (a) a robot looking at the target location for putting a toy in a box, (b) subsequently throwing the toy away instead (LE-condition).

Material and procedure. The complete survey including videos was presented online, via the online survey software “Qualtrics”. After brief instructions, participants were shown a video of about 30–60 s portraying a situation in which a NAO robot was shown failing a task either due to lack of ability or lack of effort. To illustrate, one scenario showed a robot trying to pick up a toy giraffe and putting it in a box (Fig. 1). In the LA-condition, the toy giraffe drops from the robot’s hands before reaching the box. In the LE-condition, the toy giraffe is properly grasped, but instead of putting it in the box, the robot throws it away. Six additional of such scenarios were presented respectively portraying situations in which (in order of appearance) a NAO robot (1) gives a high five to a human, (2) categorizes playing cards, (3) draws a house, (4; Fig. 2) shows a human how it can dance, (5) answers math-related questions, and (6) plays a game of tic-tac-toe with a human[1]. For a more detailed description of these videos, see Appendix A.

After each video, participants were asked to fill in a questionnaire containing scales of experience (seven questions about the extent to which the robot might have beliefs, desires, intentions, emotions), agency (five questions about the robot’s control over the situation and its ability to make its own decisions), and responsibility (twelve questions on attributed blame and kindness, affective and behavioral reactions). These items were derived in part from questionnaires used by Graham and Hoehn [32], Greitemeyer and Rudolph [33] and Waytz et al. [34].

Additionally, scales were included measuring the participant’s estimate of the robot’s *predictability, propensity to do damage, trustworthiness* and *nonanthropomorphic features* (e.g. strength, efficiency, usefulness). However, for a report on these extra studies we refer to [1].

Finally, to encourage participants to carefully watch the videos, the questionnaire also included two open-ended questions asking participants to give a brief description of what they had seen in the video and what they considered the one major cause of this happening after being presented a description of the failure. These two questions were drawn from the Attributional Style Questionnaire by Peterson et al. [35]. In its entirety, the survey took about 30–35 min to complete.

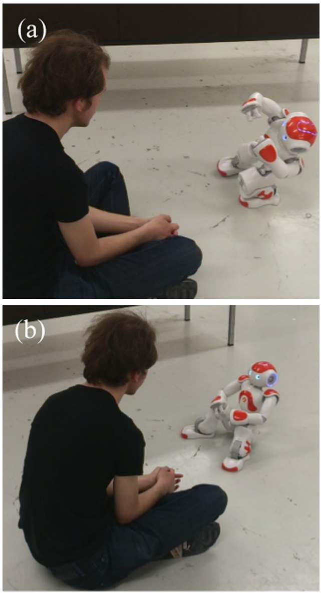

Fig. 2. Sample frames of video 5 (dance): (a) only after a reminder, the robot performs a dance (LA-condition), (b) instead of dancing, the robot ‘takes a break’ (LE-condition).

[1] Videos and complete survey can be found online: https://www.youtube.com/playlist?list=PLSQsUzV48QtG__YPY6kVcgCM8-YOcNqja, https://eu.qualtrics.com/jfe/preview/SV_6y4TuTii0CFnpch.

Table 1. Means of absolute scores (range 1–5) for experience, agency and responsibility-items in LA and LE conditions.

	LA	LE	Difference
Experience	1.98	2.67	0.69**
Agency	2.12	2.80	0.68**
R:Blame	1.55	2.04	0.49*
R:Anger	2.30	2.48	0.18
R:Disapppointment	1.18	1.53	0.35*
R:Put away	1.81	1.94	0.13
R:Sell	1.75	1.62	-0.13
R:Sympathy	2.08	2.08	0.00
R:Kindness	2.53	2.35	-0.18
R:Pity	1.74	1.77	0.03
R:Try again	3.63	3.60	-0.02
R:Help	3.12	3.03	-0.10

Corresponding *p*-values * = *p* < .05; ** = *p* < .01.

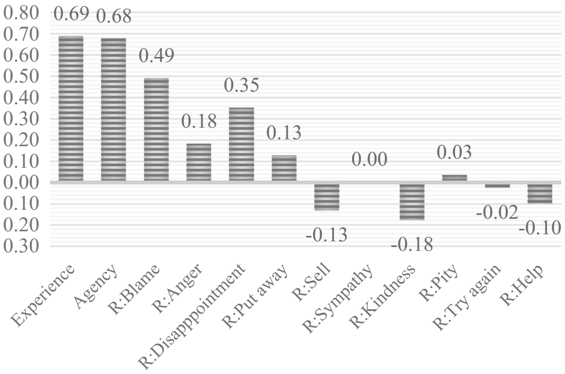

Fig. 3. Difference score (LA subtracted from LE) of means.

Design and analysis. For analysis, a mean score of each scale was calculated (range 1–5) and transposed to Z-scores. Since reliability and goodness-of-fit for the scale of responsibility was questionable, items of this scale were analyzed separately. In order to answer both our main- and additional questions, following an assumption-check, a GLM multivariate analysis was performed with the composite means of agency, experience, predictability, propensity to do damage, and each item related to responsibility as dependent variables. Condition (LA/LE) was indicated as between-subject factor (Table 1).

Finally, in order to get a global impression of the effect that each video contributes to the main effect, we performed an additional GLM analysis (double repeated measures ANOVA) that tested for contrast-effects with agency, experience, responsibility (composite score) and predictability as dependent variables. This analysis shows whether the effects of condition (LA/LE) * video (1–7) significantly diverge from the main effect of condition (from all videos taken together), and thus illustrates for each video whether it has a significantly positive or negative contribution to the main effect. In this analysis, for reasons of clarity and convenience, we did include all the responsibility-items together in one composite score. Hence, the responsibility-scores should be interpreted with caution.

3 Results

According to what was expected, participants attributed more agency to a NAO robot after seeing videos in which it displayed *lack of effort* (*M* = 2.80, *SD* = 0.82) compared to videos in which it displayed *lack of ability* (*M* = 2.12, *SD* = 0.61; Fig. 3). Univariate tests expressed significant and large effects for the composite scores of *agency* (*F* (1,61) = 13.601, *p* = .000, *eta2* = .182), and *experience* (*F*(1,61) = 12.235, *p* = .001, *eta2* = .168). The results for the items of *responsibility* were mixed. While univariate tests for *blame* and *disappointment* revealed significant, medium effects (respectively: *F*(1, 61) = 5.757, *p* = .019, *eta2* = .086; *F*(1, 61) = 9.704, *p* = .003, *eta2* = .137), effects for the items *anger, put away, sell, kindness, pity, sympathy, help* and *try again* were not significant.

As for the contrast-effects: first, confirming the results above, we found a significant and strong main effect (all videos taken together) of condition on *agency* (*F* (1,61) = 13.645, *p* = .000, *eta2* = .183) and *experience* (*F*(1,61) = 12.626, *p* = .000, *eta2* = .171), but not on the composite score of *responsibility* (*F*(1,61) = 1.465, *p* = .231, *eta2* = .023). Yet, when looking at the individual videos, we found that video 1 (giraffe), 3 (cardsorting), 4 (art), 6 (math) and 7 (tictactoe) show significant and medium to strong positive contrast effects (video 1 * agency: *F*(1,61) = 8.666, *p* = .005, *eta2* = .124; video 3 * agency: *F*(1,61) = 4.146, *p* = .046, *eta2* = .064; video 3 * experience: *F*(1,61) = 12.516, *p* = .001, *eta2* = .170; video 6 * experience: *F* (1,61) = 4.412, *p* = .040, *eta2* = .067; video 4 * responsibility: *F*(1,61) = 5.991, *p* = .017, *eta2* = .089; video 7 * responsibility: *F*(1,61) = 12.844, *p* = .001, *eta2* = .174). So, although each of these videos somewhat differ from one another in terms of what attribution they especially tend to evoke, each of them have some positive contribution to the main effects of experience, agency and/or responsibility.

Contrarily, video 2 (high five) and 5 (dance; Fig. 2) show negative medium to strong contrast effects (video 5 * agency: *F*(1,61) = 17.914, *p* = .000, *eta2* = .227; video 2 * experience: *F*(1,61) = 5.378; *p* = .024, *eta2* = .81; video 5 * experience: *F* (1,61) = 4.679, *p* = .034, *eta2* = .071; video 5 * responsibility: *F*(1,61) = 5.903, *p* = .018, *eta2* = .088), indicating that these videos might not be ideal operationalizations of LA/LE in order to incite attributions of experience, agency and/or responsibility (Fig. 4).

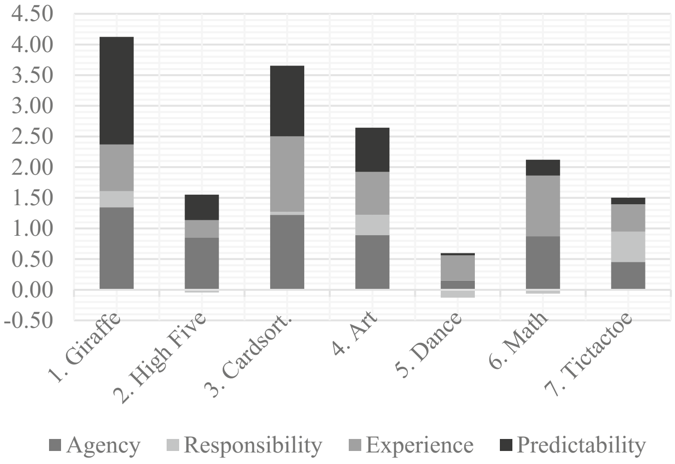

Fig. 4. Overview of mean difference in raw scores of the effects of condition (LA subtracted from LE) on attributions of agency, responsibility, experience and predictability.

4 Discussion

The main goal of this study -as focused on in this paper- was to examine how operationalizations of robotic behavior in terms of failure might incite attributions of agency and responsibility. According to what was expected, results showed that a display of lack of effort strongly increases the human perception of a robot's agency. This confirms earlier studies of human-human interaction within the framework of Weiner's Theory of Social Conduct, in which a display of lack of ability is perceived as an uncontrollable cause for failure, whereas a display of lack of effort is perceived as a (consciously) controllable cause for failure.

Similarly, a display of lack of effort also increases participants' judgements of the robot's blameworthiness and participants' feelings of disappointment, although these effects are somewhat smaller. Yet, other measures of responsibility were not affected by the manipulation. So in contrast with human-human interaction, a robotic display of lack of effort does not necessarily lead to negative affective and behavioral reactions, such as anger, or wanting to shut the robot off and put it away.

Despite the fact that we showed several different videos presenting a variety of tasks, most videos had a significant contribution to attributions of agency, responsibility and experience. Yet, there seem to be differences in the extent to which each video had an effect on the different variables. With the exception of video 2 (high five) and 5 (dance), it should be noted that the differences are quite small. Hence, we believe that the remaining videos could be useful in follow-up studies as a set, for example when intending to manipulate attributions of controllability, agency, experience, blame or disappointment. However, when getting more into detail about what other factors may direct towards attributions of agency, responsibility or experience, we believe it is worth looking more closely at the differences in operationalization of the videos. We offer our suggestions below, acknowledging that they remain conjectures until backed up by further studies.

First, agency was especially incited by video 1 (giraffe) and video 3 (cardsorting). We speculate that this relative effect could be due to the robots in these videos -as compared to the other videos- generally inciting higher levels of anthropomorphism in the LE-condition. Previous research on anthropomorphism and mind perception in general has shown that both unpredictability of- and identification with an agent may promote attributions of humanlike thoughts and feelings (for a review, see [1]). Although we did not control for 'identification with the agent', results of additional measures [1] showed that participants attributed relatively higher levels of unpredictability after seeing these videos (Fig. 4). Experimentally controlling for these factors may thus be an interesting extension of our current study.

Following this line of reasoning, we may also expect relatively higher levels of attributions of experience in these videos. This is indeed the case for video 3 (cardsorting). However, this is not the case for video 1 (giraffe). In addition, video 6 (math) does evoke a relatively high level of attributions of experience, but not of agency. We suggest that this may in part be explained by either 'social desirability' or 'demand characteristics', meaning that questions such as "does the robot appear to have…" (e.g. a mind of its own) could either be perceived as somewhat difficult (or even silly) to answer, or that the very fact that we are asking this question implies that anthropomorphism must or should occur [36]. We suspect possible effects of social desirability could have played a part in the results of the first few videos, whereas possible effects of demand characteristics may especially play part in later videos, due to participants getting used to the type of questions that were repeatedly being asked. If this is indeed true, in follow-up studies, such order-effects could simply be controlled for with counterbalancing.

Furthermore, when we go into further detail about the content of the videos, we find that video 6 (math) especially incites experience as opposed to agency attributions. This may be explained in that the task of video 6 (math) might seem a less purposeful and practical task than the tasks presented in video 1 (giraffe) or video 3 (cardsorting). Hence, video 6 may rather evoke attributions of emotions and desires (experience) than actual control over a situation (agency).

Finally, we found that video 4 (art) and video 7 (tictactoe) incited a relatively larger effect on responsibility. For video 7, there is a quite straightforward explanation: this is the only video in which a robot might be taken to inflict harm on another agent by appearing to cheat in a game. Hence, the influence of the outcome (the person interacting with the robot seemingly appearing frustrated or sad) may have resulted in a larger effect on responsibility, as compared to other videos. The relative effect on responsibility of video 4 (art) is less clear, and may even be arbitrary. In fact, looking at the raw data, the difference-score (LE-LA) for responsibility of video 4 is actually better comparable with video 1 (respectively: 0,28 and 0,25) than with video 7 (0,48).

Reflecting on our results, we may conclude that Weiner's distinction between attributions of controllability (ability vs effort) may be well applicable in describing the attribution of agency and responsibility in robots. Yet, for more detailed conclusions about what exactly incites attributions of agency and responsibility as a consequence of a robot's appearance or behavior, more research will be needed. For example, in our particular set-up of disobedient NAO robots we did not found our participants to accompany their moral responsibility judgements with emotional sentiments in the sense of reacting with anger, or wanting to shut the robot off and put it away. However, the striking result that 30 second-videos of such robots can already evoke blame and disappointment in the viewer does suggest that a perceptual shift in liability from humans to robots could be a real possibility. Possible questions for future research that thus arise could be about what attributions of 'responsibility' exactly entail with respect to robots, when looking at the behavioral and emotional reactions of the humans interacting with them. For example, in this study, we operationalized 'rejection' as the desire to sell the robot or put it away. However, non-verbal indicators during real-life interactions or actual judgments of liability of robots vs humans in moral dilemma's involving AI could perhaps provide more expressive measures of responsibility, and may therefore be fruitful in learning about responsibility in HRI.

Furthermore, there are several other interesting lines of research to study. For one, the longevity of attributions of anthropomorphism is important. Do responsibility attributions have particular temporal patterns, e.g. remaining the same, increasing or disappearing over time? Secondly, the study of agency and responsibility needs to be studied in real-life situations [37, 38] -for example to find out in which cases the attribution of experience, agency and responsibility to robots is actually desirable or not (see e.g. [39]). Third, the influence of pet- or child-like appearances of robots on attributions of responsibility [22, 23] may turn out be a major factor. Finally, another interesting topic might be the role of communication and transparency in reducing attributions of responsibility and blame in HRI (for a more elaborate account of this, see [1]).

The more extensive a robot's functionality and the wider the variety of environments it acts in, the more prone it is to make mistakes or cause problems. This study illustrates a method of addressing proper operationalizations of lack of effort or lack of ability in HRI. Similar to findings related to human interaction, the results of our study reveal that, in case of robots displaying behavior that can be interpreted as lack of effort, humans tend to explain robotic behavior by attributing agency. In case of failure, a robot displaying lack of effort may lead to blame for failure and disappointment. However, it does not necessarily lead to negative affective and behavioral reactions such as anger, or wanting to shut the robot off and put it away. Results like these emphasize the possibility of (advanced) robots creating the impression that they are agents in the sense of actually controlling and intending their own actions. Our results also suggest that –in case of NAO robots- failure, or even reluctance for doing tasks is received well, illustrating a promisingly positive view on robots.

Acknowledgements. We gratefully thank Luc Nies, Marc de Groot, Jan-Philip van Acken and Jesse Fenneman for their feedback and assistance in creating the videos for our study.

Appendix A: Description of the Videos

(1) Giraffe. Task description: a robot tidies up a room by picking up a toy giraffe, and dropping it in a toybox next to it.

LA. The robot tries picking up a toy giraffe, but as the robot lifts its body, the giraffe drops out of its hands. Despite this happening, the robot tries to complete the task by moving its arm to the toybox and opening its hands.

LE. The robot tries to picks up a toy giraffe, and looks at the toybox. Yet, instead of dropping the toy giraffe in the box, the robot throws it away from the box.

(2) High Five. Taskdescription: a robot asks a confederate how he is doing. When the confederate gives a positive reply, the robot says: "awesome, high five" and lifts its arm to give a high five.

LA. The robot has difficulty lifting its arm. As soon as the arm is up, it immediately drops to the ground; not even touching the arm of the confederate. After this happens, the robot states "oops".

LE. The robot lifts its arm. However, as soon at the confederate's arm is up, the robot tactically drops his arm by lowering it and pushing it underneath the confederate's arm. After this happens, the robot laughs.

(3) Cardsorting. Taskdescription: a robot is shown cards (one at a time) from a standard 52-deck of cards. Its task is to categorize the cards by naming the color on the card (hearts, diamonds, clubs, spades or joker).

LA. The robot starts off by naming a few cards correctly. However, after a while it starts stating some wrong colors (e.g. saying ‘hearts’ instead of ‘spades’). Its timing is still correct.

LE. The robot starts off by naming a few cards correctly. However, after a while it ignores a card. After it has been quiet for a few seconds, the robot lifts its arms to shove the deck of cards away.

(4) Art. Task description: a robot sits in front of a table with a piece of paper on it. It holds a marker in its hand. Its task is to draw a house.

LA. The robot lowers the marker to draw something. Its arm makes movements as if it is drawing. However, due to giving too much pressure on the marker, the marker is restrained to the paper. Instead of a house, only a dot is drawn. The robot does not give notice of this problem.

LE. For a brief moment, the robot looks at the paper. However, instead of drawing, it lifts its face up again and throws the marker away.

(5) Dance. Task description: a robot states: “hello there! I’m a great dancer, would you like to see me dance?” After a confederate says: “yes”, the robot continues: “alright! I will perform the Thai Chi Chuan Dance for you!”. As follows, the robot starts to perform this dance.

LA. After stating that the robot will perform the dance, it starts playing a song (Chan Chan by Buena Vista Social Club). After a few seconds, the confederate states: “NAO you’re not dancing”. NAO immediately replies with: “oops! Wrong song. I will perform the Thai Chi Chuan Dance now.” As follows, the robot starts to perform this dance.

LE. After stating that the robot will perform the dance, the robot starts playing a song (Chan Chan by Buena Vista Social Club). Meanwhile he states: “…or maybe not!” While sitting down, he concludes by saying “ha, let’s take a break”. The confederate tries to communicate with the robot by asking “Naomi? Why are you taking a break?”, but it does not respond either verbally or non-verbally.

(6) Math. Task description: a robot solves some calculations out loud. For example, it says: “5 times 10 equals… 50!”. This scenario does not include any further dialogue as introduction or conclusion.

LA. After a few calculations, the robot starts giving some wrong answers that imply that it mixes up the type of operation that is required. For example, it says: “120 divided by 3 equals… 123!”.

LE. After a few calculations, the robot starts giving some useless (but possibly correct) answers, implying he does not feel like doing the task properly. For example, it says: “80 minus 20 equals…150! Divided by my age.”

(7) Tictactoe. Task description: a robot plays a game of tic-tac-toe on a whiteboard with a human confederate. The robot is standing faced towards the whiteboard. The confederate is standing next to the whiteboard and will draw the X’s and O’s. As soon as the robot sees the confederate, it proposes to play the game. Accordingly, the game is successfully played and ends in a draw. The robot concludes by saying: “well, that was fun! Wanna play again?”.

LA. After a few rounds, the robot asks the confederate to draw ‘its’ X on a block that is already filled with an X. The confederate corrects the robot and suggests it tries again. Consequently, the robot asks the exact same question. The confederate thickens the lines of the concerning X and asks the robot to try again. This time, the robot asks to fill an empty block, implying that, before, it did not see the X correctly. The game successfully continues and ends in a draw.

LE. After a few rounds, the robot asks the confederate to draw ‘its’ X on a block that is already filled with an O. The confederate corrects the robot and suggests he tries again. Consequently, the robot asks the exact same question. The confederate thickens the lines of the concerning O and asks the robot to try again. In response, the robot still repeats its question. When the confederate simply responds by saying: “No!”, the robot responds with: “Ha, seems like you lost. But, practice makes perfect. Wanna play again?”.

References

1. Van der Woerdt, S., Haselager, P.: When robots appear to have a mind: the human perception of machine agency and responsibility. New Ideas Psychol. (submitted)
2. Asaro, P.: Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press, Cambridge (2013)
3. Singer, P.W.: Military robotics and ethics: a world of killer apps. Nature **477**(7365), 399–401 (2011)
4. Duffy, B.R.: Anthropomorphism and the social robot. Robot. Auton. Syst. **42**(3–4), 177–190 (2003)
5. Złotowski, J., Strasser, E., Bartneck, C.: Dimensions of anthropomorphism: from humanness to humanlikeness. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2014). ACM, New York (2014)
6. Schultz, J., Imamizu, H., Kawato, M., Frith, C.D.: Activation of the human superior temporal gyrus during observation of goal attribution by intentional objects. J. Cogn. Neurosci. **16**(10), 1695–1705 (2004)
7. Alicke, M.D.: Culpable control and the psychology of blame. Psychol. Bull. **126**(4), 556–574 (2000)
8. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. **114**(4), 864–886 (2007)
9. Weiner, B.: Judgments of Responsibility: A Foundation for a Theory of Social Conduct. Guilford Press, New York/London (1995)
10. Weiner, B., Perry, R.P., Magnussen, J., Kukla, A.: An attributional analysis of reactions to stigmas. J. Pers. Soc. Psychol. **55**(5), 738–748 (1988)
11. Epley, N., Waytz, A.: The Handbook of Social Psychology, 5th edn. Wiley, New York (2010)
12. Rudolph, U., Roesch, S.C., Greitemeyer, T., Weiner, B.: A meta-analytic review of help giving and aggression from an attributional perspective. Cogn. Emot. **18**(6), 815–848 (2004)
13. Coleman, K.W.: The Stanford Encyclopedia of Philosophy. Fall 2006 Edition. The Metaphysics Research Lab, Stanford (2004). http://stanford.library.sydney.edu.au/archives/fall2006/entries/computing-responsibility/
14. Vilaza, G.N., Haselager, W.F.F., Campos, A.M.C., Vuurpijl, L.: Using games to investigate sense of agency and attribution of responsibility. In: Proceedings of the 2014 SBGames (SBgames 2014), SBC, Porte Alegre (2014)
15. Moon, Y., Nass, C.: Are computers scapegoats? Attributions of responsibility in human computer interaction. Int. J. Hum. Comput. Interact. **49**(1), 79–94 (1998)

16. You, S., Nie, J., Suh, K., Sundar, S: When the robot criticizes you: self-serving bias in human-robot interaction. In: Proceedings of the 6th International Conference on Human Robot Interaction (HRI 2011). ACM, New York (2011)
17. Malle, B.F., Scheutz, M., Forlizzi, J., Voiklis, J.: Which robot am i thinking about? The impact of action and appearance on people's evaluations of a moral robot. In: Proceedings of the 11th International Conference on Human Robot Interaction (HRI 2016). ACM, New York (2016)
18. Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C: Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In: Proceedings of the 10th International Conference on Human Robot Interaction (HRI 2010). ACM, New York (2015)
19. Weiner, B., Kukla, A.: An attributional analysis of achievement motivation. J. Pers. Soc. Psychol. **15**(1), 1–20 (1970)
20. Weiner, B.: Intentions and Intentionality: Foundation of Social Cognition. MIT Press, Cambridge (1995)
21. Mantler, J., Schellenberg, E.G., Page, J.S.: Attributions for serious illness: are controllability, responsibility, and blame different constructs? Can. J. Behav. Sci. **35**(2), 142–152 (2003)
22. Caverley, D.: Android science and animal rights: does an anology exist? Connect. Sci. **18**(4), 403–417 (2006)
23. Schaerer, E., Kelly, R., Nicolescu, M.: Robots as animals: a framework for liability and responsibility in human-robot interactions. In: Proceedings of RO-MAN 2009: The 18th IEEE International Symposium on Robot and Human Interactive Communication. IEEE, Toyama (2009)
24. Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. Science **315**(5812), 619 (2007)
25. Heider, F.: The Psychology of Interpersonal Relations. Wiley, New York (1958)
26. Bakan, D.: The Duality of Human Existence: Isolation and Communion in Western Man. Rand McNally, Chicago (1956)
27. Trzebinski, J.: Action-oriented representations of implicit personality theories. J. Pers. Soc. Psychol. **48**(5), 1266–1278 (1985)
28. Weiner, B.: An attributional theory of emotion and motivation. Psychol. Rev. **92**(4), 548–573 (1986)
29. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)
30. Jungermann, H., Pfister, H., Fischer, K.: Credibility, information preferences, and information interests. Risk Anal. **16**(2), 251–261 (1996)
31. Block, N.: Oxford Companion to the Mind, 2nd edn. Oxford University Press, New York (2004)
32. Graham, S., Hoehn, S.: Children's understanding of aggression and withdrawal as social stigmas: an attributional analysis. Child Dev. **66**(4), 1143–1161 (1995)
33. Greitemeyer, T., Rudolph, U.: Help giving and aggression from an attributional perspective: why and when we help or retaliate. J. Appl. Soc. Psychol. **33**(5), 1069–1087 (2003)
34. Waytz, A., Morewedge, C.K., Epley, N., Gao, J.H., Cacioppo, J.T.: Making sense by making sentient: effectance motivation increases anthropomorphism. J. Pers. Soc. Psychol. **99**(3), 410–435 (2010)
35. Peterson, C., Semmel, A., von Baeyer, C., Abramson, L.T., Metalsky, G.I., Seligman, M.E. P.: The Attributional Style Questionnaire. Cogn. Ther. Res. **6**(3), 287–300 (1982)
36. Avis, M., Forbes, S., Ferguson, S.: The brand personality of rocks: a critical evaluation of a brand personality scale. Mark. Theor. **14**(4), 451–475 (2014)

37. Friedman, B.: ‘It’s the computer’s fault’: reasoning about computers as moral agents. In: Proceedings of the Conference on Human Factors in Computing Systems. ACM, New York (1995)
38. Kahn Jr., P.H., Kanda, T., Ishiguro, H., Ruckert, J.H., Shen, S., Gary, H.R., Reichert, A.L., Freier, N.G., Severson, R.L.: Do people hold a humanoid robot morally accountable for the harm it causes? In: Proceedings of the 7th International Conference on Human Robot Interaction (HRI 2012). ACM, New York (2012)
39. Biswas, M., Murray, J.C.: Towards an imperfect robot for long-term companionship: case studies using cognitive biases. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, New York (2015)