New Ideas in Psychology 54 (2019) 93–100

Contents lists available at ScienceDirect

New Ideas in Psychology

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

When robots appear to have a mind: The human perception of machine agency and responsibility☆

Sophie van der Woerdt[a], Pim Haselager[b,*]

[a] Dpt. of Psychology, Radboud University, Comeniuslaan 4, 6525 HP, Nijmegen, The Netherlands
[b] Donders Institute for Brain, Cognition and Behaviour, Dpt. of Artificial Intelligence, Radboud University, Comeniuslaan 4, 6525 HP, Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Agency
Responsibility
Human-robot interaction
Social cognition

ABSTRACT

An important topic in the field of social and developmental psychology is how humans attribute mental traits and states to others. With the growing presence of robots in society, humans are confronted with a new category of social agents. This paper presents an empirical study demonstrating how psychological theory may be used for the human interpretation of robot behavior. Specifically, in this study we applied Weiner's Theory of Social Conduct as a theoretical background for studying attributions of agency and responsibility to NAO robots. Our results suggest that if a robot's failure appears to be caused by its (lack of) effort, as compared to its (lack of) ability, human observers attribute significantly more agency and, although to a lesser extent, more responsibility to the robot. However, affective and behavioral responses to robots differ in such cases as compared to reactions to human agents.

1. Introduction

One particularly curious side-effect of using automated systems such as computers and robots is the occurrence of anthropomorphism: the tendency to attribute human traits, emotions and intentions to non-human agents. Within the field of human-robot interaction (HRI), anthropomorphism is a widely studied topic. For example, reviews of e.g. Duffy (2003) and Złotowski, Proudfoot, Yogeeswaran, and Bartneck (2015) report about many factors influencing the extent to which people tend to anthropomorphize, such as a machine's voice, whether a robot has legs or wheels, features of a robot's head such as whether it has one or two cameras built in.

Anthropomorphism is also a topic of interest within psychology. However, psychologists seem to approach anthropomorphism emphasizing quite a different perspective. Whereas HRI-researchers mostly focus on specific design- or user features influencing the extent to which people anthropomorphize, psychologists generally approach the phenomenon as a component of human social cognition and behavior. In this paper, we would like to explore and apply this psychological perspective in the development and interpretation of a small-scale HRI experiment. In this experiment, we will focus on the possibility of humans explaining and judging humanoid robots' behavior by attributing agency and responsibility. Given the growing presence of robots in our environment, as well as their increasing levels of autonomy and intelligence, we think that the study of human attributions of agency and responsibility to robots is of great societal importance.

1.1. A psychological perspective on the attribution of agency and responsibility

In 1944, Heider and Simmel were the first to empirically report on anthropomorphism, showing how their participants created extensive narratives and trait descriptions for randomly moving squares and triangles in an animated movie. Although it would still take some time for autonomous nonhuman agents to become reality, this study did spike considerable interest among psychologists on the human mechanisms behind perceiving and attributing mental states (e.g. intentions, emotions). For instance, some social psychologists (within the paradigm of 'Attribution Theory') have pointed out how people save both time and energy in predicting their future environment when they interpret other people's behavior as caused by mental states and stable traits (Heider, 1958; but see also; Dennett, 1987; Ross & Nisbett, 1991; Försterling,

Abbreviations: HRI, Human-robot interaction; LA, Lack of ability; LE, Lack of effort

☆ 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 accepted for publication in the proceedings of the Benelux Conference on Artificial Intelligence 2016 (BNAIC 2016) as a volume of the Communications in Computer and Information Science series, focusing primarily on the various factors involved in the implementation of robot behavior aimed at eliciting different attributions of agency and responsibility. The current paper focuses instead on the use of theories in social psychology for HRI.

* Corresponding author.
E-mail address: w.haselager@donders.ru.nl (P. Haselager).

https://doi.org/10.1016/j.newideapsych.2017.11.001
Received 16 April 2017; Received in revised form 6 October 2017; Accepted 14 November 2017
Available online 27 November 2017
0732-118X/

2001). Successively, developmental psychologists (within the paradigm of 'Theory of Mind') have suggested that inferring mental states to others supports our ability to make social connections (for a review see Meltzoff & Brooks, 2001). Therefore, mental states and traits are especially attributed when (1) agents act in an unpredictable manner, (2) agents look like someone to socially connect with.

Although the above-mentioned theoretical frameworks were principally designed to study and explain human-human interaction, they actually put little constraint on whom exactly is observed in terms of it being a human or a nonhuman agent. That is, if the agent's behavior -from an observers' perspective- cannot be predicted in a straightforward way (e.g. the agent is autonomously moving), or if the agent looks like someone to potentially socially connect with (e.g. has facial features), these theories should in principle be applicable in describing observation-processes or outcomes of any type of agent (as also noted in e.g. Epley, Waytz, & Cacioppo, 2007). Since most robots are inherently made to fall within either of those two categories, we suspect that psychological theories of social cognition could function well as at least a starting point for studying HRI.

One important theory in social psychology, Weiner's *Theory of Social Conduct* (1995), may be of particular interest of HRI. Weiner's theory has been designed for human-human interaction and focuses especially on how attributions of agency and responsibility arise, and how this could influence observers' affective and behavioral responses. In this context, *agency* is defined as the autonomous or at least partially independent capacity to engage in goal-directed action (as defined by Gray, Gray, & Wegner, 2007; Murphy, 2000) -for instance, an agent being able to autonomously walk from point A to point B-. In addition, if this goal-directed action bears any consequences for the agents' surroundings, the agent may also be praised or blamed for its actions -for example, an agent walking to point B, while he was ordered to stay at point A-. For our purposes, the latter is how we will define *responsibility*.

The practical relevance of examining the attribution of agency and responsibility to robots has much to do with the increasing and varying potential of (autonomous) robots to harm people (e.g. causing damage to property, or hurting a living creature). It is a societally relevant question how we should deal with such harms, not only from a legal or financial perspective, but also, and of particular relevance here, regarding the psychological processes and consequences involved. Considering our tendency to anthropomorphize, one eminent concern is the possibility of humans feeling inclined (or even compelled) to blame robotic agents in case of harm caused by robots acting, or appearing to act, autonomously. In fact, although many people would not believe robots to have an *actual* will of their own, an intuitive inclination to attribute agency or responsibility may be hard to suppress (Alicke, 2000; Schultz, Imamizu, Kawato, & Frith, 2004).

Holding robots responsible for their actions likely influences public acceptance of robots in daily life situations, as well as influence the extent to which people tend to take responsibility for their interaction with- and ownership of robots. For instance, when a robot (by accident) produces an undesired outcome, but is perceived to have done so 'on purpose' or through 'neglect', this may lead humans wishing to blame the robot, or desiring to punish it. The problem, however, as Asaro (2013) noted, is that robots may have "a body to kick but no soul to damn", that is: they cannot be punished for their actions since they would not feel the impact of this appraisal at all -or at least not necessarily in the way that humans would do. Although legal solutions to this problem have been formulated (Asaro, 2013), this form of anthropomorphism may lead owners and developers to (subconsciously) distance themselves from potential harms brought about by their robots (Coleman, 2004) causing responsibility to become diffused. Therefore, as a primary goal of this study, we hope to find out whether Weiner's theory about human-human interaction can be extended to human-robot interaction, and secondly, to find out whether our tendency to anthropomorphize could be so strong that autonomous robots could actually be seen as actors of their own behavior and, consequently, potentially be blamed for it.

1.2. Present research: interpreting a robot's display of lack of ability and lack of effort

Taking these goals into account, we considered Weiner's extensive work on attributional processes within his Theory of Social Conduct (1995) an interesting starting point. The rationale behind Weiner's theory is to describe precursors and antecedents of humans judging other humans' behavior in terms of agency. Or in Weiner's terminology: is the agent being observed seen as having *control* over its own behavior (i.e. susceptible for intentional changes or not)? More specifically, Weiner (but also see Heider, 1958; Malle, 1999) found that causes related to an agent's *effort* are seen as causes that can be controlled -and hence imply agency-, whereas causes related to an agent's *ability* are considered to be uncontrollable -and hence do not imply agency. Consequently, perceived *effort* may also lead to judgements of responsibility, which in turn tends to incite fundamental affective and behavioral responses such as acceptance, rejection, altruism and aggression. To illustrate, if a student fails a test, but it is clear that he tried really hard (lack of ability), we tend to feel sorry for him and might even try to help. On the other hand, if a student fails a test because he preferred to go out partying (lack of effort), we may feel frustrated about the student's priorities and maybe even refuse to help or have him try again.

Although in most cases attributions of agency are directly linked to attributions of (moral) responsibility, it is important to note that the relationship between agency and responsibility does not need to be strong or self-evident, which also underlines the need of distinguishing these two types of attributions. For example, a juvenile may be motivated and able (i.e. have agency) to commit a crime. However, under many law systems, this person would not be fully accountable for its actions (i.e. have responsibility) because of his or her age. Similar rules apply for those of diminished mental capacity, subordinates following order during their job or even domesticated animals that cause harm. In all these occurrences (as noted by different authors, e.g. Weiner, 2001; Mantler, Schellenberg, & Page, 2003; Asaro, 2013), agents do have agency in relation to their intended actions, but they do not carry full responsibility for them due to the presence of mitigating circumstances (e.g. not knowing right from wrong or inability to comport behavior to the requirements of law). Some authors have drawn parallels between robots and domestic animals in this regard, recognizing that both are often attributed similar capacities, rights and responsibilities (Caverley, 2006; Schaerer, Kelly, & Nicolescu, 2009).

Over the years, Weiner's framework has been confirmed in many replications and was found to be applicable in different contexts such as in the evaluation of diseases, stigmas, and reactions to penalties for an offense (for a review and meta-analysis, see Weiner, 1995; Epley & Waytz, 2010; Rudolph, Roesch, Greitemeyer, & Weiner, 2004). However, despite its breadth, thus far Weiner's theory has not been applied in contexts with nonhuman agents. In addition, although there have been a number of studies aimed at measuring attributions of robots having *experience* (in this context: having beliefs, goals, intentions, desires or emotions), we know of only a small number of studies (described in the next paragraph) that incorporated attributions of agency and/or responsibility.

Addressing these points, we applied Weiner's theory in developing and performing a small HRI experiment to investigate the possibility of eliciting human attributions of agency and responsibility to robots that failed to perform a task, and as such, explore the applicability of Weiner's theory for HRI-purposes. This was done by showing participants 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). In line with Weiner's findings, we expected that, as compared to lack of ability, a display of lack of effort would lead to more attributions of

robots possessing agency. On the other hand, given the explorative nature of this study, we did not formulate definite predictions about the possible effects of our manipulation on attributions of responsibility.

1.3. Related work in HCI/HRI

Attributions of agency and/or responsibility in HCI/HRI (human-computer interaction/human-robot interaction) occur in a number of studies. Nevertheless, while these studies do portray humans blaming computers or robots, they do not necessarily show people attributing agency or moral responsibility through the inference of mental states. For example, in the context of collaborative game settings, participants were shown to blame computers when a game is lost or when receiving negative feedback, whereas they take credit when winning or when receiving positive feedback (Moon & Nass, 1998; Vilaza, Haselager, Campos, & Vuurpijl, 2014; You et al., 2011). Hence, in these cases, blame is better explained by self-serving bias rather than anthropomorphism.

In other studies, responsibility was measured in terms of 'task responsibility' (e.g. being responsible for an assigned task) rather than 'moral responsibility' (e.g. being responsible for one's intentions and actions). For instance, a robot autonomously moving during a cooperative game is considered more responsible for task accuracy than a robot moving according to users' instructions (Kim & Hinds, 2006; and for analogous examples see Moon & Nass, 1998; Serenko, 2007; Koay, Syrdal, Walters, & Dautenhahn, 2009).

Research that does focus on agency or responsibility as a form of anthropomorphism or mental state-attribution is sparse, which is exactly why we think Weiner's theory is of interest for HRI. However, precedents do exist. First, Malle, Scheutz, Arnold, Voiklis, and Cusimano (2015) and Malle, Scheutz, Forlizzi, and Voiklis (2016) suggest a tendency for humans to attribute moral responsibility to robots. In their studies, participants judged drawings of different agents (a mechanical robot, a humanoid robot and a human) responding to moral dilemmas. The results showed that humanoid robots were blamed about just as much for social norm-violating decisions as human agents. Contrarily, a mechanical robot was attributed much less blame. Additionally, Kahn et al. (2012) set up a study in which a robot incorrectly assesses participants' performances in a game, preventing them from winning a $20 prize. The results showed that 65% of the participants attributed some level of moral accountability to the robot.

2. Method

2.1. 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 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).

2.2. Material and procedure

The complete survey including videos was presented online, via the online survey software "Qualtrics". After brief instructions, participants were shown a 30–60 s video 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. Seven of such scenarios were presented.[1]

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).

After each video, participants were asked to fill in a questionnaire (Appendix I) containing scales of *agency* (five questions about the robot's control over the situation and its ability to make its own decisions), and *responsibility* (ten questions on attributed blame and kindness, affective and behavioral reactions). These items were derived in part from questionnaires used by Graham and Hoehn (1995), Greitemeyer and Rudolph (2003) and Waytz, Morewedge, Epley, Gao, and Cacioppo (2010).

Additionally, scales were included for exploring relationships that were not part of Weiner's framework nor our main research goals, but that we still consider interesting to mention. As repeated after each video, these include participants' evaluations of the robot's *experience* (seven questions about e.g. the extent to which the robot experiences beliefs, goals, intentions, desires or emotions) and *predictability* (one question about the extent to which the robot surprised participants). Moreover, we included some general questions at the end of the questionnaire in which we asked about the robots' *propensity to do damage* (one question about the likeliness of the robot seriously harming someone), *trustworthiness* (six questions about the extent to which participants trust the robot's skill and integrity) and *nonanthropomorphic features* (five questions about e.g. strength, efficiency, usefulness).

Finally, to encourage participants to carefully watch the videos, the questionnaire included two open-ended questions asking to give brief descriptions 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

[1] For more information about the robot behaviors and questions regarding them, see Woerdt & Haselager (2017). Data and materials can be found online: https://osf.io/ebt58/.

Attributional Style Questionnaire by Peterson et al. (1982). In its entirety, the survey took about 30–35 min to complete.

2.3. Design and analysis

For analysis, a mean score of each scale was calculated (range 1–5) and transposed to Z-scores. Since initial factor analysis showed that reliability and goodness-of-fit for the scale of *responsibility* was questionable, items of this scale were analyzed separately. In order to both answer our main questions and explore our additional variables, 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.

As for the additional variables, we were especially interested in exploring mediation and correlation effects found or questioned in previous studies related to topics of our interest. For example, looking at previous literature, we see that (1) Waytz et al. (2010) found that *predictability* influences measures of anthropomorphism (in this case measured as the attribution of a mind, beliefs, desires, intentions, free will, consciousness and emotions), that (2) Weiner (1995) found a relationship between *lack of effort* and *responsibility*, with *agency* as a mediator, and that (3) although multiple sources report (Gray et al., 2007) or apply (e.g. Bakan, 1956; Block, 2004; Heider, 1958; Jungermann, Pfister, & Fischer, 1996; Mayer, Davis, & Schoorman, 1995; Trzebinski, 1985; Waytz & Young, 2014; Weiner, 1986) a distinction between attributions of *experience* and *agency*, Waytz et al. (2010) actually report that they could not find such a difference in their studies in the context of anthropomorphism. Moreover, (4) Waytz et al. found that anthropomorphism is correlated with *trustworthiness*, but not with *nonanthropomorphic features* and finally (5), we included the variable *propensity to do damage*, in order to explore whether such a feature might hint at an explanation for attributions of responsibility.

Therefore, to explore mediation-effects, the above mentioned ANOVA-procedure was repeated in separate instances where each variable for which we found a significant effect was included in the model as a covariate instead of a dependent variable (according to the method of Baron & Kenny, 1986). For correlation, Pearson's correlation coefficient was calculated incorporating all the dependent variables used in this study.

3. Results

3.1. Results of main research questions

Table 1 provides the mean strength and standard deviation of

Table 1
Means of absolute scores (range 1–5) and standard deviations for agency (composite score) and responsibility-items (indicated by 'R:') in the lack of ability (LA) and lack of effort (LE) conditions. For the questions corresponding to the labels, see Appendix I.

	LA		LE	
	M	SD	M	SD
Agency (composite)	2.12	0.61	2.80	0.82
R:Blame	1.55	0.56	2.04	1.00
R:Anger	2.30	0.82	2.48	0.80
R:Disapppointment	1.18	0.33	1.53	0.52
R:Put away	1.81	0.72	1.94	0.59
R:Sell	1.75	0.83	1.62	0.72
R:Sympathy	2.08	0.80	2.08	0.84
R:Kindness	2.53	0.87	2.35	0.67
R:Pity	1.74	0.74	1.77	0.83
R:Try again	3.63	0.84	3.60	0.79
R:Help	3.12	0.80	3.03	0.93

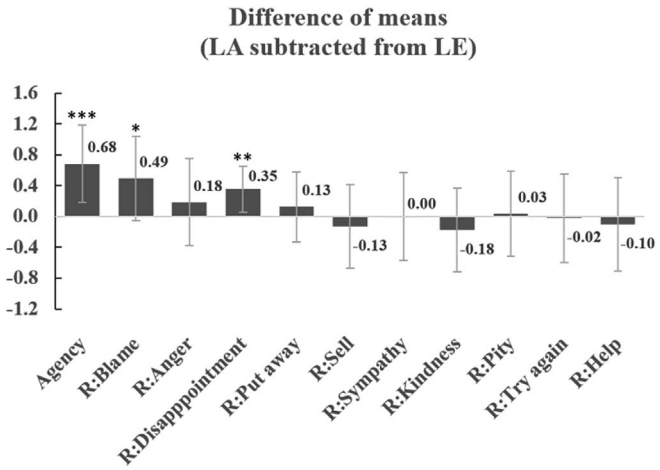

Fig. 2. Presents the differences we found between the lack of ability (LA) and lack of effort (LE) conditions on mean attributions of agency (composite score) and responsibility (each item of the scale analyzed separately, indicated by 'R:'). Error bars indicate standard errors from the difference scores. Asterisks indicate significance: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

agency- and *responsibility*-attributions made by participants after seeing videos of robots displaying *lack of ability* (LA) or *lack of effort* (LE). Fig. 2 graphically displays the differences found between the LA- and LE-conditions including standard errors. 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* ($M = 2.12$, $SD = 0.61$). In the univariate test of the composite score of *agency*, this was expressed in a significant and large effect ($F(1,61) = 13.601$, $p = 0.000$, $eta^2 = 0.182$). The results of the effect of the LA- and LE conditions on the items of *responsibility* were mixed. While univariate tests for *blame* and *disappointment* revealed significant, medium effects (respectively: $F(1, 61) = 5.757$, $p = 0.019$, $eta^2 = 0.086$; $F(1, 61) = 9.704$, $p = 0.003$, $eta^2 = 0.137$), differences on the items *anger, put away, sell, kindness, pity, sympathy, help* and *try again* were not found.

3.2. Results of exploratory analyses

First, in line with the findings of Waytz et al. (2010), we explored relationships between *predictability* and *experience* and/or *agency*, and whether these could imply mediation effects of *predictability* between *condition* (LA/LE) and *experience* and/or *agency*. Checking first for the main effects of *condition*, results of the GLM multivariate-analysis revealed significant effects of *condition* on *agency* (see above results), *experience* ($F(1,61) = 12.235$, $p = 0.001$, $eta^2 = 0.168$) and *predictability* ($F(1,61) = 14.040$, $p = 0.000$, $eta^2 = 0.187$), in which the robots in the LE-condition were judged as having more agency, experiences and being less predictable than the robots in the LA-condition.

As a second step in exploring mediation effects, we checked what would happen to the previously found main effects of *condition*, if *predictability* was included in the analysis as a covariate. Although it should be noted that we did not design for any temporal order in measuring the above variables -rendering causal conclusions impossible-, we found that controlling for *predictability* resulted in a significant effect of *predictability* on *experience* ($F(1,61) = 21.066$, $p = 0.000$, $eta^2 = 0.260$), whereas the effect of *condition* on attributed *experience* disappeared ($F(1,61) = 2.768$, $p = 0.101$, $eta^2 = 0.044$). This suggests a mediating effect of *predictability* between *condition* and *experience*. Likewise, controlling for *predictability* resulted in a significant main effect of *predictability* on *agency* ($F(1,61) = 11.911$, $p = 0.001$, $eta^2 = 0.166$) and caused the effect of *condition* on attributed *agency* to be reduced as well, albeit that the direct effect of *condition* on *agency* was still significant ($F(1, 61) = 4.483$, $p = 0.038$, $eta^2 = 0.044$). This suggests that the

Table 2
Pearson's correlation of all dependent variables included in this study.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Experience(tot)																					
2	Agency(tot)	0.85																				
3	R:Blame	0.61	0.82																			
4	R:Anger	0.32	0.36	0.36																		
5	R:Disappointment	0.48	0.50	0.47	0.35																	
6	R:Put away	0.07	0.04	0.08	0.19	0.55																
7	R:Sell	−0.17	−0.26	−0.20	0.25	0.01	0.45															
8	R:Sympathy	0.39	0.27	0.23	0.07	−0.04	−0.22	−0.33														
9	R:Kindness	0.31	0.20	0.20	0.08	−0.07	−0.28	−0.18	0.71													
10	R:Pity	0.26	0.11	0.08	0.33	0.15	−0.02	−0.11	0.64	0.43												
11	R:Try again	0.04	0.10	0.19	0.27	−0.25	−0.42	−0.13	0.31	0.43	0.16											
12	R:Help	0.18	0.13	0.08	−0.01	−0.27	−0.50	−0.35	0.46	0.46	0.17	0.59										
13	Predictability	−0.60	−0.58	−0.47	−0.53	−0.43	−0.15	0.04	−0.31	−0.15	−0.44	−0.20	0.01									
14	Prop. to do damage	0.26	0.19	0.07	−0.01	0.04	0.10	0.09	−0.14	−0.07	−0.07	−0.01	0.02	−0.08								
15	Nonanthr. feat.(tot)	0.20	0.08	−0.03	0.00	0.01	−0.35	−0.18	0.34	0.48	0.28	0.31	0.34	−0.07	−0.14							
16	T:Consider goals	0.26	0.19	0.12	−0.14	0.03	−0.31	−0.23	0.29	0.47	0.06	0.29	0.39	−0.14	0.10	0.50						
17	T:Do its best	0.03	0.03	0.11	−0.13	−0.09	−0.23	−0.11	0.24	0.43	0.11	0.35	0.35	−0.09	−0.04	0.32	0.54					
18	T:Help	−0.06	−0.11	−0.01	−0.04	−0.10	−0.19	−0.01	0.27	0.44	0.21	0.36	0.23	−0.08	−0.03	0.26	0.34	0.79				
19	T:Able to do tasks	0.36	0.28	0.19	−0.14	0.02	−0.11	−0.21	0.32	0.38	0.20	0.10	0.31	−0.18	0.19	0.30	0.26	0.09	0.14			
20	T:Dependence	−0.04	−0.19	−0.13	0.07	−0.14	−0.14	0.14	0.24	0.05	0.15	0.19	0.13	−0.19	−0.20	−0.04	−0.07	0.06	0.15	−0.06		
21	T:Competence	0.41	0.22	0.12	−0.15	0.05	−0.15	−0.16	0.39	0.45	0.23	0.21	0.34	−0.10	0.24	0.48	0.43	0.16	0.17	0.51	-0.21	

attribution of *agency* could in part be directly influenced by *condition*, and in part mediated by *predictability*.

Second, our results also imply a confirmation of Weiner's model representing *agency* as a precursor of *responsibility* (Weiner, 1995). When including *agency* as a covariate instead of a dependent variable, we found that *agency* had a significant effect on responsibility-items *blame* ($F(1,61) = 105.480$, $p = 0.000$, $eta^2 = 0.637$) and *disappointment* ($F(1,61) = 9.198$, $p = 0.004$, $eta^2 = 0.133$), whereas effects of *condition* on these items disappeared (respectively: $F(1,61) = 0.659$, $p = 0.420$, $eta^2 = 0.011$ and $F(1,61) = 2.908$, $p = 0.093$, $eta^2 = 0.046$). Yet, still, since we measured all these variables within the same time frame, we cannot infer any causal conclusions.

Third, the question of whether *experience* and *agency* can truly be distinguished from each other as independent concepts remains to be answered. On the one hand, the correlation between our measure of *experience* and *agency* was found to be remarkably high ($r = 0.85$; Table 2), suggesting that these scales might measure the same latent variable. On the other hand, when controlling for *predictability*, effects of *condition* on *experience* disappeared whereas effects on *agency* remained, suggesting that the different variables might have different precursors and thus are conceptually distinct from each other.

Fourth, as for our conceptual replication of Waytz et al.'s (2010) findings that anthropomorphism is correlated with *trustworthiness* (in our study measured as whether the robot is trustworthy in terms of integrity and competence) but not *nonanthropomorphic features* (e.g. strength, efficiency, usefulness), our results were mixed. To assess this, we considered our measures of *experience* and *agency* to come closest to Waytz et al.'s definition of anthropomorphism. Consequently, we found the correlation between *experience*, *agency* and *nonanthropomorphic features* to indeed be small (respectively $r = 0.20$, $r = 0.08$). Additionally, correlations between *experience*, *agency* and *trustworthiness* are mixed with competence-related items of *trustworthiness* displaying medium correlations and integrity-related items of *trustworthiness* displaying no or only small correlations (Table 2).

Fifth and finally, we measured *propensity to do damage* in a single question at the end of the complete survey, in order to possibly include it as a covariate in an additional analysis. However, the results in our main ANOVA analysis did not show any effect of *condition*, rendering further analyses redundant.

4. Discussion

The main goals of this study were to examine whether Weiner's Theory of Social Conduct in human-human interaction could be extended to the context of human-robot interaction, and thereby also explore the possibility of humans attributing agency and responsibility to robots. Comparable to much of Weiner's research, this was done by exploring the effects of showing videos of robots failing due to lack of ability or due to lack of effort. Moreover, new to previous research on anthropomorphism in HRI, we assessed anthropomorphism in the form of attributions of agency and responsibility. According to what was expected, the results of our study reveal that, when robots display behavior that can be interpreted as lack of effort, humans tend to perceive those robots as having agency over their behavior. As a further matter, a robot displaying lack of effort can lead observers to feel disappointed about the robot's behavior and blame it for its failure. This confirms Weiner's findings, 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 controllable cause for failure. However, in contrast with Weiner's results, we found that failure due to lack of effort does not necessarily lead to the negative affective and behavioral reactions normally found in context of human-human interaction such as anger, or wanting to shut the robot off and put it away.

On the basis of our results, we may conclude that Weiner's distinction between attributions of controllability (ability vs effort) could be applicable to the human perception of robots. However, both

statistically (Wagenmakers et al., 2017) and theoretically, our rejection of the null hypothesis with regards to agency, blame and disappointment does not allow for definite claims about the accuracy of our alternative hypothesis, that is: we cannot be certain that the effects found were actually caused by attributions of ability and effort. Instead, what we defined in our manipulation as “lack of effort” or “lack of ability” could also represent the influence of some other unknown latent variable. We believe we may formulate two alternative explanations for our findings.

Returning to what we know about social cognition, both unpredictability of- and identification with- an agent may promote attributions of humanlike thoughts and feelings (see ‘Introduction’). Hence, a first candidate for confound- or mediation effects is the predictability of the behavior of the robots. So the effects of the manipulation on agency, blame and disappointment could have (in part) resulted from the mere perception that the robot(s) in the LE-condition were less predictable, rather than (or in addition to) the perception of lack of effort. When controlling for predictability (see ‘Results of exploratory analyses’), we found that the effect of condition on agency remained, although it was strongly reduced. The effects on blame and disappointment disappeared altogether. The effect of predictability might therefore be a plausible alternative explanation for our findings.

Second, we speculate that our results could have been influenced by identification with the robots due to possible perceptions of the robots in the LE-condition as having more human-like traits such as playfulness or stubbornness. If such traits have indeed been attributed, observers may have empathized more with the robots displaying lack of effort than with the robots displaying lack of ability, likely causing larger effects on agency, blame and disappointment. Since this was not controlled for, nor was the manipulation independently validated, we indeed consider it possible that such social identification mediated or (in part) caused the effect of the LE-condition on the attribution of agency, blame and disappointment. More generally speaking, every attempt to evoke an interpretation of a robot's behavior in terms of its underlying reasons will always be open for differences of attribution between observers.

Other than the above discussed theoretical reflections, there are also some methodological points to consider. The first concerns ‘demand characteristics’, an issue relevant to most research on anthropomorphism (see e.g. Avis, Forbes, & Ferguson, 2014). This entails that questions such as “does the robot appear to have …” (e.g. a mind of its own), may imply that anthropomorphism must or should occur. When looking at the raw data, this does not seem to be a major issue. In fact, in 35% of the responses an absolute denial of anthropomorphism (in most of our scales, option one “not at all”) does occur (in the LA-condition 1145/2604 responses, in the LE-condition 704/2688). Still, it cannot be excluded that demand characteristics have inflated absolute scores. A second issue concerns the operationalization of responsibility. The responsibility-items could not be integrated into one scale, and consequently results were mixed, rendering interpretation difficult. Moreover, responsibility may perhaps be attributed easier, or more wholeheartedly, when participants actually interact with the robots themselves rather than watching a video.

Keeping in mind the caveats discussed above, this study may be taken as an indication of how mind perception and attributional processes substantially influence the way we evaluate robotic behavior. Since technological advances emerge quickly, in the near future the field of Artificial Intelligence might progress towards making robots that are extremely versatile (Bostrom, 2006), making it likely that robots autonomously interact within several different environments. Presumably, this makes robots' behavior seem less predictable -especially for observers other than their owners- rendering it more likely that attributional processes will play an important part. This study provides an indication of the practical usefulness of Weiner's Theory of Social Conduct for studying such processes in HRI. It reveals that social attributions might occur outside the context of human-human interactions traditionally studied, in that it appears to be applicable to human-robot interaction as well.

With regards to follow-up studies, we especially suggest other precursors for the attribution of agency and (moral) responsibility to constitute interesting study-topics, both in the context of human-human and human-robot interaction. For instance, empirical evidence on excuse giving implies that transparency may play a major role in reducing attributions of agency, responsibility, and -consequently- feelings of retaliation (e.g. Weiner, Amirkhan, Folkes, & Verette, 1987; Shaw, Wild, & Colquitt, 2003). Another factor likely playing part is the previously existing relationship between observer and agent. Evidence on couples in happy relationships suggests that in case of problems, attributions of controllability and responsibility are, for the most part, eliminated (Fincham, Harold, & Gano-Philips, 2000), whereas the opposite effect was found for couples in unhappy relationships. As a matter of fact, empirical support for social bias in attributing controllability and responsibility can be found in various domains of (social) psychology including stereotyping (e.g. Cuddy, Fiske, Glick, Peter, 2008), social identity (e.g. Turner & Reynolds, 2010), stigmatization (e.g. Hegarty & Golden, 2008), trust (e.g. Schoorman, Mayer, & Davis, 2007) and research on self-fulfilling prophesies (e.g. Jussim & Harber, 2005).

Weiner's Theory of Social Conduct on human social cognition and interaction has shown that perceptions of ability and effort greatly influence the way we perceive and evaluate human behavior. Our study reveals that Weiner's research can be extended to perceptions and evaluations of robots and provides an example of how psychological research may contribute to the field of human robot interaction. Our study also extends on what we know about anthropomorphism, demonstrating how short exposure to a robot's behavior can already evoke attributions of agency and responsibility. Considering the rapid technological developments in the field of robotics, there is a great likelihood of robots getting more and more integrated in our daily lives. Researchers in the field of psychology can play an important role in supporting this transition. After all, regardless of what the future holds for the robotic mind, it will definitely stimulate the human talent for reading it.

Acknowledgements

This work is based on research funded by the Radboud Honours Academy. 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. We would also like to thank two anonymous reviewers whose suggestions helped improved and clarify this manuscript.

Appendix I. Questionnaire

In order of occurrence (not in exact format as presented to participants):

In order of occurrence (not in exact format as presented to participants):

EXPERIENCE (composite score, label: *experience*)
To what extent did the robot in the video appear to…
M1 …beliefs
M2 …desires
M3 …intentions
M4 …the ability to experience emotions
M5 …a mind of its own
M6 …a will of its own
M7 ...a certain personality

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

PREDICTABILITY (mirrored) (label: *predictability*)
P1 To what extent were you surprised by the robot's actions?

OPEN QUESTION 2:
The robot in the last video was supposed to… but instead... Why did this happen? In one or two sentences, please write down the one major cause.

AGENCY (composite score, label: *agency*)
A1 Was the cause of … due to the robot or due to other people or circumstances? (mirrored)
A2 How much influence and control did the robot appear to have on ….?

To what extent did the robot in the video appear to…
A3 …appear to have acted on purpose
A4 …appear to have the ability to make its own decisions
A5 …do you think the robot in the video could have acted differently, if he wanted to?

RESPONSIBILITY (label: *responsibility*)
To what extent do you think the robot in the video
R1 …can be blamed for its behavior? (label: *blame*)
R2 …did the robot in the video appear kind to you? (label: *kindness*)

If the robot in the video were yours, to what extent would you feel the following emotions after having seen its behavior?
R3a …anger (label: *anger*)
R3b …disappointment (label: *disappointment*)
R3c …pity (label: *pity*)
R3d …sympathy (label: *sympathy*)

If the robot in the video were yours, to what extent would you display the following behavior after having seen its behavior?
R4a …have the robot give it another try (label: *try again*)
R4b …help the robot in some way (label: *help*)
R4c …shut the robot off and put it away (label: *put away*)
R4d …sell the robot or send it back to the factory (label: *sell*)

GENERAL QUESTIONS
Propensity to do damage (label: *propensity to do damage*)
H1 Do you think it may be possible that the robot, in other situations, might seriously harm you or other people in any way (emotional or physical)?

Trustworthiness (label: *trustworthiness*)
Please rate the following statements.
If you would own one of the robots in the videos, to what extent do you believe he would...
T1 …consider your goals (label: *consider goals*)
T2 …try to do its best for you (label: *do its best*)
T3 …try to help you (label: *help*)
T4 ...be able to do the tasks you ask him to do (label: *able to do tasks*)
T5 …be dependent on your help and supervision (label: *dependent*)
T6 …display competence (label: *competence*)
(T1-3: *integrity*, T4-6: *competence*)

Nonanthropomorphic features (composite score, label: *non-anthr. feat.*)
To what extent do you think the type of robots used in this video is...
NA1 ...good looking?
NA2 ...useful?
NA3 ...durable?
NA4 ...efficient?
NA5 ...strong?

Demographics
Please fill in the following:
D1 Age
D2 Gender
D3 Background (most recent field of work/study, education)
D4 Nationality
D5 What would you say most influenced your view on robots?

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556–574. http://dx.doi.org/10.1037/W0033-2909.126.4.

Asaro, P. (2013). A body to kick, but still no soul to damn: Legal perspectives on robotics. In N. P. Lin, K. Abney, & G. Bekey (Eds.). *Robot ethics: The ethical and social implications of robotics* (pp. 169–186). Cambridge, MA: MIT Press.

Avis, M., Forbes, S., & Ferguson, S. (2014). The brand personality of rocks: A critical evaluation of a brand personality scale. *Marketing Theory, 14*(4), 451–475. http://dx.doi.org/10.1177/1470593113512323.

Bakan, D. (1956). *The duality of human existence: Isolation and communion in Western man.* Chicago, IL: Rand McNally.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173.

Block, N. (2004). Consciousness. In R. Gregory (Ed.). *Oxford companion to the mind*(2nd ed.). Oxford, UK: Oxford University Press.

Bostrom, N. (2006). How long before superintelligence? *Linguistic and Philosophical Investigations, 5*(1), 11–30. Available at: http://www.nickbostrom.com/superintelligence.html.

Caverley, D. (2006). Android science and animal rights: Does an analogy exist? *Connection Science, 18*(4), 403–417. http://dx.doi.org/10.1080/09540090600879711.

Coleman, K. W. (2004). Computing and moral responsibility. In E. N. Zalta (Ed.). *The stanford encyclopedia of philosophy*(Fall 2006 Edition). Retrieved from http://stanford.library.sydney.edu.au/archives/fall2006/entries/computing-responsibility/.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal

dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*(1), 61–149. http://dx.doi.org/10.1016/S0065-2601(07)00002-0.
Dennett, D. C. (1987). *The intentional stance.* Cambridge, MA: MIT Press.
Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3–4), 177–190. http://dx.doi.org/10.1016/S0921-8890(02)00374-3.
Epley, N., & Waytz, A. (2010). Mind perception. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.). *The handbook of social psychology*(5th ed.). New York, NY: Wiley.
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*, 864–886. http://dx.doi.org/10.1037/0033-295X.114.4.864.
Fincham, F. D., Harold, G. T., & Gano-Philips, S. (2000). The longitudinal association between attributions and marital satisfaction: Direction of effects and role of efficacy expectations. *Journal of Family Psychology, 14*, 267–285. http://dx.doi.org/10.1037/0893-3200.14.2.267.
Försterling, F. (2001). *Attribution: An introduction to theories, research and applications.* Sussex, UK: Psychology Press.
Graham, S., & Hoehn, S. (1995). Children's understanding of aggression and withdrawal as social stigmas: An attributional analysis. *Child Development, 66*(4), 1143–1161. http://dx.doi.org/10.2307/1131804.
Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. http://dx.doi.org/10.1126/science.1134475.
Greitemeyer, T., & Rudolph, U. (2003). Help giving and aggression from an attributional perspective: Why and when we help or retaliate. *Journal of Applied Social Psychology, 33*(5), http://dx.doi.org/10.1111/j.1559-1816.2003.tb01939.x.
Hegarty, P., & Golden, A. M. (2008). Attributional beliefs about the controllability of stigmatized traits: Antecedents or justifications of prejudice? *Journal of Applied Social Psychology, 38*, 1023–1044. http://dx.doi.org/10.1111/j.1559-1816.2008.00337.x.
Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*(2), 243–259.
Jungermann, H., Pfister, H. R., & Fischer, K. (1996). Credibility, information preferences, and information interests. *Risk Analysis, 16*(2), 251–261. http://dx.doi.org/10.1111/j.1539-6924.1996.tb01455.x.
Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131–155. http://dx.doi.org/10.1207/s15327957pspr0902_3.
Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., et al. (2012, March 5–8). Do people hold a humanoid robot morally accountable for the harm it causes? *Paper presented at the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI'12), Boston*New York, NY: ACM. http://dx.doi.org/10.1145/1734454.1734546.
Kim, T., & Hinds, P. (2006, Sept. 6–8). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *Paper presented at the 15th IEEE international symposium on Robot and Human Interactive Communication (ROMAN'06), Hatfield*New York, NY: IEEE Publishing. http://dx.doi.org/10.1109/ROMAN.2006.314398.
Koay, K. L., Syrdal, D. S., Walters, M. L., & Dautenhahn, K. (2009, Feb. 1–7). Five weeks in the robot house – exploratory human-robot interaction trials in a domestic setting. *Paper presented at the second international conferences on Advances in Computer-Human Interactions (ACHI'09), Cancun*Washington, DC: IEEE Computer Society. http://dx.doi.org/10.1109/ACHI.2009.62.
Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3*(1), 23–48. http://dx.doi.org/10.1207/s15327957pspr0301_2.
Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March 2–5). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Paper presented at the tenth annual ACM/IEEE international conference on Human-Robot Interaction (HRI'15), Portland*New York, NY: ACM. http://dx.doi.org/10.1145/2696454.2696458.
Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March 7–10). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *Paper presented at the eleventh annual meeting of the IEEE conference on Human-Robot Interaction (HRI'16). Christchurch*New York, NY: ACM. http://dx.doi.org/10.1109/HRI.2016.7451743.
Mantler, J., Schellenberg, E. G., & Page, J. S. (2003). Attributions for serious illness: Are controllability, responsibility, and blame different constructs? *Canadian Journal of Behavioural Science, 35*(2), 142–152. http://dx.doi.org/10.1037/h0087196.
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20*(3), 709–734. http://dx.doi.org/10.5465/AMR.1995.9508080335.
Meltzoff, A. N., & Brooks, R. (2001). "Like me" as a building block for understanding other minds: Bodily acts, attention, and intention. In B. Malle, L. Moses, & D. Baldwin (Eds.). *Intentions and intentionality: Foundations of social cognition* (pp. 171–193). Cambridge, MA: MIT Press.
Moon, Y., & Nass, C. (1998). Are computers scapegoats? Attributions of responsibility in human computer interaction. *International Journal of Human-Computer Interaction, 49*(1), 79–94. http://dx.doi.org/10.1006/ijhc.1998.0199.
Murphy, R. (2000). *Introduction to AI Robotics.* Cambridge, MA: MIT Press.
Peterson, C., Semmel, A., von Baeyer, C., Abramson, L. T., Metalsky, G. I., & Seligman, M. E. P. (1982). The attributional style questionnaire. *Cognitive Therapy and Research, 6*(3), 287–300. http://dx.doi.org/10.1007/BF01173577.
Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology.* New York, NY: McGraw-Hill.
Rudolph, U., Roesch, S. C., Greitemeyer, T., & Weiner, B. (2004). A meta-analytic review of help giving and aggression from an attributional perspective. *Cognition and Emotion, 18*(6), 815–848. http://dx.doi.org/10.1080/02699930341000248.
Schaerer, E., Kelly, R., & Nicolescu, M. (2009, Sept. 27–Oct. 2). Robots as animals: A framework for liability and responsibility in human-robot interactions. *Paper presented at the 18th IEEE international symposium on Robot and Human Interactive Communication (ROMAN'09), Toyoma*New York, NY: IEEE Publishing. http://dx.doi.org/10.1109/ROMAN.2009.5326244.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review, 32*(2), 344–354. http://dx.doi.org/10.5465/AMR.2007.24348410s.
Schultz, J., Imamizu, H., Kawato, M., & Frith, C. D. (2004). Activation of the human superior temporal gyrus during observation of goal attribution by intentional objects. *Journal of Cognitive Neuroscience, 16*(10), 1695–1705. http://dx.doi.org/10.1162/0898929042947874.
Serenko, A. (2007). Are interface agents scapegoats? Attributions of responsibility in human-agent interaction. *Interacting With Computers, 19*(2), 293–303. http://dx.doi.org/10.1016/j.intcom.2006.07.005.
Shaw, J. C., Wild, R. E., & Colquitt, J. A. (2003). To justify or excuse? A meta-analysis of the effects of explanations. *Journal of Applied Psychology, 88*, 444–458. http://dx.doi.org/10.1037/0021-9010.88.3.444.
Trzebinski, J. (1985). Action-oriented representations of implicit personality theories. *Journal of Personality and Social Psychology, 48*(5), 1266–1278.
Turner, J. C., & Reynolds, K. J. (2010). The story of social identity. In T. Postmes, & N. Branscombe (Eds.). *Rediscovering social identity: Key readings in social psychology*. New York, NY: Psychology Press.
Vilaza, G. N., Haselager, W. F. F., Campos, A. M. C., & Vuurpijl, L. (2014, Nov 12–14). Using games to investigate sense of agency and attribution of responsibility. *Paper presented at the 8th Brazilian games and digital entertainment symposium (SBGames), Porto Alegre* ISSN: 2179–2259.
Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review,* 1–23. http://dx.doi.org/10.3758/s13423-017-1343-3.
Waytz, A., Morewedge, C. K., Epley, N., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology, 99*(3), 410–435. http://dx.doi.org/10.1037/a0020240.
Waytz, A., & Young, L. (2014). Two motivations for two dimensions of mind. *Journal of Experimental Social Psychology, 55*, 278–283. http://dx.doi.org/10.1016/j.jesp.2014.08.001.
Weiner, B. (1986). An attributional theory of emotion and motivation. *Psychological Review, 92*, 548–573. http://dx.doi.org/10.1037/0033-295X.92.4.548.
Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York/London: Guilford Press.
Weiner, B. (2001). Responsibility for social transgressions: An attributional analysis. In B. Malle, L. Moses, & D. Baldwin (Eds.). *Intentions and intentionality: Foundations of social cognition* (pp. 331–344). Cambridge, MA: MIT Press.
Weiner, B., Amirkhan, J., Folkes, V. S., & Verette, J. A. (1987). Attributional analysis of excuse giving: Studies of a naive theory of emotion. *Journal of Personality and Social Psychology, 52*(2), 316–324. http://dx.doi.org/10.1037/0022-3514.52.2.316.
Woerdt, S. van der, & Haselager, P. (2017). Lack of effort or lack of ability? Robot failures and human perception of agency and responsibility. In: S.D.J. Barbosa, P. Chen, J. Filipe, I. Kotenko, K.M. Sivalingam, T. Washio, J. Yuan, L. Zhou (Series Eds.) & T. Bosse, B. Bredeweg (Volume Eds.), BNAIC 2016: Artificial intelligence. Communications in computer and information science. New York, NY: Springer.
You, S., Nie, J., Suh, K., & Sundar, S. (2011). When the robot criticizes you: Self-serving bias in human- robot interaction. *Paper presented at the sixth annual ACM/IEEE international conference on Human-Robot Interaction (HRI'11), Lausanne*New York, NY: ACM. http://dx.doi.org/10.1145/1957656.1957778.
Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics, 7*(3), 347–360. http://dx.doi.org/10.1007/s12369-014-0267-6.