International Journal of Social Robotics (2022) 14:1137–1153
https://doi.org/10.1007/s12369-021-00856-9

The Fundamental Attribution Error in Human-Robot Interaction: An Experimental Investigation on Attributing Responsibility to a Social Robot for Its Pre-Programmed Behavior

Aike C. Horstmann[1] · **Nicole C. Krämer**[1]

Accepted: 1 December 2021 / Published online: 16 January 2022

Abstract
Since social robots are rapidly advancing and thus increasingly entering people's everyday environments, interactions with robots also progress. For these interactions to be designed and executed successfully, this study considers insights of attribution theory to explore the circumstances under which people attribute responsibility for the robot's actions to the robot. In an experimental online study with a $2 \times 2 \times 2$ between-subjects design ($N = 394$), people read a vignette describing the social robot Pepper either as an assistant or a competitor and its feedback, which was either positive or negative during a subsequently executed quiz, to be generated autonomously by the robot or to be pre-programmed by programmers. Results showed that feedback believed to be autonomous leads to more attributed agency, responsibility, and competence to the robot than feedback believed to be pre-programmed. Moreover, the more agency is ascribed to the robot, the better the evaluation of its sociability and the interaction with it. However, only the valence of the feedback affects the evaluation of the robot's sociability and the interaction with it directly, which points to the occurrence of a fundamental attribution error.

Keywords Human-robot interaction · Agency · Autonomy · Attribution theory · Fundamental attribution error · Humanoid robot · Experimental study

The ethics committee of the division of Computer Science and Applied Cognitive Sciences at the Faculty of Engineering of the University of Duisburg-Essen approved the study and written informed consent was obtained. The fundamental attribution error in human-robot interaction: An experimental investigation on attributing responsibility to a social robot for its pre-programmed behavior.

1 Introduction

Social robots are rapidly evolving and with the advancements of this technology, they are increasingly taking over tasks for which interacting with humans is an essential necessity. Since social robots are progressively operating in people's everyday environments such as homes, work places, schools, hospitals, and museums, human-robot interactions are becoming more socially situated and multi-faceted [1]. With these social, multi-facetted interactions the question arises how people interpret the robots' actions, i.e., whether and to what extent they attribute responsibility to a robot for its actions. According to attribution theory, people constantly attempt to understand the causes and implications of another person's behavior [2, 3]. Based on Kelley's covariation model, different factors determine whether (problematic) behavior is attributed to an internal or an external cause [3]. However, people tend to overestimate the influence of a person's disposition while neglecting situational factors [2]. This is the starting point for this study, which aims to examine the effects of feedback by the social robot Pepper that is either believed to be autonomous or pre-programmed. In addition, different circumstances such as the valence of the feedback as well as the robot's expected future role are considered.

Various studies show that people automatically apply and react to polite or even flattering behavior when interacting with a computer [4–6]. Particularly because those interactive

This manuscript features online supplementary material (OSM), which can be accessed here: https://osf.io/6 × 7rj/.

✉ Aike C. Horstmann
aike.horstmann@uni-due.de

[1] Social Psychology: Media and Communication, University of Duisburg-Essen, Duisburg, Germany

devices take over the role of an assistant or service provider, positive feedback is expected and negative feedback leads to a worsened evaluation of the device [7, 8]. The study's focus is to examine whether a robot is evaluated differently when there a high external justification for the robot's negative feedback, i.e., if the robot will be evaluated more positively if a fixed program is believed to generate negative feedback opposed to the robot being believed to create the feedback autonomously by itself.

People's feelings towards social robots are often mixed due to the prevalence of two very different prospects for them [9]. On the one hand, people fear that robots will become competitors to them, mainly professionally but also in their private lives [10–12]. With this scenario in mind, people rather prefer robots to be as controllable as possible and not to act autonomously. In contrast to this negative view, social robots are also portrayed and expected to make people's lives easier by functioning as helpful assistants in domestic, public, or work environments [11, 13, 14]. Focusing on this prospect, people rather prefer autonomous robots to reduce the workload and enhance comfort and convenience. Therefore, we assume that the expected future role of the robot affects how its level of autonomy is perceived and evaluated. Since previous research showed that a social robot is evaluated differently depending on whether it is framed before as a threatening competitor or a helpful assistant [15], we also aim to examine whether the feedback-giving social robot is perceived differently when it is framed as an assistant or a competitor beforehand.

To sum up, the aim of this study is to examine whether and what kind of attribution processes occur when people interact with social robots. For this purpose, it is analyzed whether the information that the either negative or positive feedback is pre-programmed or autonomously generated by the robot affects people's evaluations of this robot and the interaction with it. Additionally, since expectations regarding robots are ambiguous but might greatly influence people's perception, this study further considers the effect of framing the robot as becoming an assistant or a competitor in the future.

1.1 Autonomous vs. Pre-Programmed Feedback

In order to shape successful future human-robot interactions, particularly collaborations, and to avoid frustrations as well as errors and conflicts, it is crucial to examine the attribution of credit and blame in those interactions [16]. According to attribution theory, people endeavor to understand what factors cause or contribute to other people's behaviors with the aim to understand and predict their future behavior better [2, 3]. Against this background, Kelley's covariation model presents three different factors, consensus, consistency, and distinctiveness, which are critical for determining whether a (problematic) behavior should be attributed to an internal or an external cause [3]. Consensus is present if the person's response is similar to other persons' responses to the same stimulus, consistency maps a person's uniform response to the stimulus over time and in different situations, and distinctiveness describes whether the person only responds to this stimulus in this way [3]. Low consensus and distinctiveness paired with a high consistency should lead to increased person attribution, high consensus, distinctiveness, as well as consistency to stimulus attribution, and high consensus paired with low distinctiveness as well as consistency to circumstance attribution [17]. The covariation model follows a logical structure in determining how people supposedly form attributions. However, empirical studies show that people pervasively tend to overestimate the influence of an observed person's disposition, i.e., internal factors, while underestimating the situational context, i.e., external factors. This was labelled as fundamental attribution error or overattribution effect [2, 18, 19]. For instance, in a study by Jones and Harris [20], people attributed chance-directed behaviors to disposition rather than to the situation. The attitude of an essay which people were asked to write was determined by a coin toss. However, people still attributed the imposed attitude as the writer's actual attitude [20].

Considering these findings from interpersonal studies leads to the question which factors, external or internal, are predominantly considered when a robot's behavior is evaluated. Thus, in this study it is evaluated whether the presence or absence of a high external justification for the robot's behavior affects how the robot is evaluated afterwards. More specifically, it is examined whether the robot is evaluated more positively when a fixed program supposedly generates the negative, unpleasant feedback opposed to when people believe the robot creates the feedback autonomously by itself.

Autonomy goes along with an internal locus of causality, placing the variables to explain a behavior within a person [21, 22]. However, an internal locus of causality requires a certain amount of agency, i.e., the capacity to act independently and free of choice [23]. Agency is a form of self-motivated governance [24] which manifests in various mechanisms such as autonomy [22], animacy [25], and free will [26]. Since agency presupposes consciousness, it is rather clearly ascribed to natural agents such as humans in contrast to artificial agents such as robots [27]. However, borders between natural and artificial entities are blurring and it may be more important to focus on whether an artificial agent is perceived as acting autonomously than whether it is able to possess an own consciousness.

From an objective point of view, machines are not able to act completely autonomously since their actions are controlled by their users and/or programmers. However, following the insights of the media equation theory [28], people may be inclined to perceive and treat a robot as if it is acting autonomously when it is described to create feedback by

itself and displays behavior that underlines this expectation. According to Bartneck and Forlizzi [29], the autonomy of a social robot represents its technological capability to act without direct input from a human and those autonomous actions are perceived as intentional. Here, we argue that when examining the impact for human-robot interaction, whether the robot is believed to act independently from external input is a critical factor. An experiment by Kim and Hinds [16] showed that people attribute more blame to a highly autonomous robot than to themselves or other participants compared to when the robot displays low autonomy. Interestingly, people shifted blame for errors but not credit for successes to the autonomous robot [16]. This is consistent with findings of attribution theory which indicate that people tend to blame others for errors and give credit to themselves for successes [2]. In a different setting, people feel less responsible when collaborating with a human-like than a machine-like robot, which could be interpreted in a sense that a human-like robot is perceived as more autonomous and thus more capable of carrying responsibility than a machine-like robot [30]. Considering these accumulated findings, the focus of this work is on the perception of autonomy, independent of whether the robot is acting autonomously from a technological point of view.

Since autonomous behavior is related to perceived agency [22], a social robot providing feedback which is believed to be generated by itself should lead to an increased sense of agency of this robot. Agency was further found to be connected to responsibility, such as deserving punishment for wrongdoing [31]. Therefore, the robot should be held more responsible for the content of the feedback when it is believed to have generated it itself compared to when the feedback is believed to be pre-programmed by someone else. Furthermore, the robot believed to be acting autonomously should also be perceived as more competent than a robot which just bluntly and mindlessly passes on what its programming tells it to.

Thus, the following is hypothesized:

H1 When a social robot is believed to give autonomous feedback, this leads to a higher perceived (a) agency of the robot, (b) responsibility of the robot for the content of the feedback, and (c) competence of the robot, compared to when the robot is believed to give pre-programmed feedback.

A social robot's believed autonomy should increase how agentic people perceive this robot since the two concepts are related [22]. Agency in turn contributes to the notion of the robot as active social agent causing people to develop strong affective and emotional bonds with robots [1, 32, 33]. Thus, a social robot which is perceived as more agentic should be evaluated more positively regarding its sociability and the interaction with this robot should be evaluated more positively compared to a robot which is perceived as less agentic.

Consequently, the following mediation hypothesis is postulated:

H2 When a social robot is believed to give autonomous feedback, this leads to a higher perceived agency of this robot, which in turn leads to a more positive evaluation of (a) the robot's sociability and (b) the interaction with the robot, compared to when the robot is believed to give pre-programmed feedback.

1.2 Valence of Feedback

In general, when people receive negative feedback, their intrinsic motivation and their own perceived competence decreases [34]. When people experience a discrepancy between their own performance and internally or externally set standards, they are motivated to reduce this discrepancy, usually by attaining the standard [35]. However, other coping strategies are for instance to change the standards, to reject the feedback which points to the discrepancy, or to escape or avoid the situation [35, 36]. People also show increased effort to evaluate others more poorly when criticized to maintain their own self-esteem [37].

Since people often react to interactive media as if they were real persons [28], the question arises how people react to negative feedback by a social robot. Particularly since most interactive devices provide a service to humans, they are expected to adhere to social norms, i.e., to be polite and friendly. According to Sayin and Krishna [38], the more human-like characteristics a device shows, the more it is expected to behave politely. Several studies with interactive computers showed that people apply politeness [4] and respond to flattery [5, 6] as if they were interacting with another person. For instance, a computer which gives polite, positive feedback, which does not even have to be sincere or contingent to people's actual performance, is evaluated significantly better compared to a computer providing neutral feedback [5, 6]. Consequently, negative, unfriendly feedback in comparison to positive, friendly feedback should pose a negative expectancy violation leading to detrimental communication and relationship outcomes [39]. In this vein, negative computer-generated feedback was found to be related to increased task response times [40], stronger persuasive effects [41], heightened response evaluation processes [42], and surprise as well as frustration [43]. Thus, it is not surprising that when artificial interlocutors give negative feedback, they are evaluated negatively [7], particularly when it is rather personal, emotional negative feedback [8].

Another aspect to take into account is reciprocity, which is considered as a fundamental norm of social interactions [44]. Research showed that subjects also display reciprocal behavior with non-human interaction partners like computers [45], robots [46, 47], and virtual agents [48–50]. Mostly, reciprocity is examined in the context of self-disclosure [49],

establishment of rapport [48], or mimicry [50]. However, reciprocity may also result in retaliation in case of negative behaviors. For instance, when people were confronted with an agent displaying risky behavior in a social game, they switched from cooperative to competitive strategies [51]. Likewise, after interacting with a tough agent in a negotiation setting, participants were subsequently more willing to use deceiving negotiation tactics such as lying and negative emotions [52]. In a study by Fogg and Nass [45], there was also evidence for retaliation effects when participants worked with a computer that was not very helpful.

Summing up, people tend to evaluate others more poorly when criticized [37], in particular technology that provides criticism [7, 8], and also tend to retaliate when treated in a negative, uncooperative way [45, 51]. Consequently, when a robot provides negative feedback, the robot's sociability and competence as well as the general interaction with it should be evaluated more negatively than when it provides positive feedback.

Against this background, the following hypothesis is postulated:

H3 When a social robot provides negative feedback, this leads to a more negative evaluation of (a) the robot's sociability and (b) the interaction with the robot, compared to when the robot provides positive feedback.

In addition to the main effects of the valence of the robot's feedback and whether it is believed to be generated autonomously or pre-programmed, there should also be interaction effects of these two factors. Due to people's deeply rooted aversion to personal rejection [53], negative feedback should lead to a negative evaluation of the robot's sociability and the general interaction with it, particularly when this feedback is believed to be generated autonomously by this robot, i.e., when the robot is perceived to have chosen those harsh words itself. However, when there is a clear external justification for the robot's unpleasant behavior, such as that it was previously programmed to respond exactly this way to its opponent's answers, people should perceive and evaluate the robot's sociability and the interaction with it not as negatively as with the supposedly autonomously acting robot.

Based on these deliberations, the following is hypothesized:

H4 When a social robot's feedback is negative and believed to be pre-programmed, this leads to a more positive evaluation of (a) the robot's sociability and (b) the interaction with the robot, compared to when the feedback is negative and believed to be autonomous.

1.3 Robots' Expected Future Roles

Looking at expected future roles of social robots, there are basically two rather contradicting prospects – one is highly desired, the other one rather met with fears and worries. Both images are portrayed prominently in media, which causes people to have double-minded feelings towards robots [9]. Due to the prominence of those two opposing views, this study aims to analyze the influence of the views separately by emphasizing one or the other.

The negative, even feared prospect places social robots in the roles of competitors. People are worried about losing control over robots. This, going to the extreme, results in fears of humans being replaced or dominated by robots and is called "Frankenstein Syndrome" [11, 12, 54]. Mass media, which are widely accessible and thus have an extensive reach, often depict on these ideas of robots developing their own consciousness and revolting against humans [10]. In a study by Horstmann and Krämer [13], the results indicate that the recall of "bad" fictional robot characters leads people to have greater fears that robots might outrace and become a threat to humans. Interacting with a social robot which is described with the aim to become better and more efficient than humans and to take away tasks from them, should be evaluated as undesirable. Consequently, this robot's sociability as well as the interaction with it should be evaluated poorly.

The other opposing view portrays social robots as helpful assistants in domestic, public, as well as work environments [13]. The prospect of having social robots take over unpleasant or strenuous tasks, which would make life considerably easier, is evaluated as highly desirable [11, 13, 14]. Interacting with a social robot which is portrayed as striving to become a valuable help with the aim to make people's everyday life more comfortable, should be highly desirable. As a result, this robot's sociability as well as the interaction with it should be evaluated positively.

A previous study examining the influence of emphasizing either one of those two contrary prospects already showed that a social robot expected to become a helpful assistant is evaluated more sociable than a robot expected to compete with humans for their jobs [15]. Based on these findings as well as the theoretical deliberations regarding the two prominent but contradicting views on social robots, it is hypothesized that portraying the robot in a negative way (describing it as a competitor working against humans), the robot's perceived sociability as well as the interaction with it are evaluated more negatively compared to when the robot is framed in a positive light (describing it as an assistant working for humans):

H5 When a social robot is expected to take over the role of a competitor, this leads to a more negative evaluation of (a) the robot's sociability and (b) the interaction with the robot, compared to when the robot is expected to become an assistant.

According to the assumptions of the expectancy violation theory [39], people may hold higher standards for a high reward person compared to a low reward person, which makes it possible for the high reward person to commit a more

serious expectancy violation. A person's reward valence represents how desirable it is to interact with this person [55]. Transferring this to human-robot interaction, a robot believed to aspire being a beneficial help should have a high reward valence, while a robot believed to strive to compete with humans for their jobs should have a low reward valence [56]. Consequently, assuming that the phenomena described by the expectancy violation theory also occur when interacting with humanoid robots, negative feedback provided by the assistant robot should cause a more severe expectancy violation, which goes in line with detrimental communication outcomes, than negative feedback provided by the competitor robot [39]. Based on these deliberations, the current work aims to analyze whether a favorable, desirable framing of a social robot leads to a more negative evaluation of its sociability and the interaction with it when it offers harsh, negative feedback compared to negative feedback coming from a robot portrayed in an unfavorable way.

Against this background, it is hypothesized:

H6 When a social robot is believed to become an assistant, the robot's negative feedback leads to a more negative evaluation of (a) the robot's sociability and (b) the interaction with the robot, compared to when the robot is believed to become a competitor.

Additionally, a three-way interaction of the three manipulation variables - valence of the social robot's feedback, believed autonomy of the robot, and expected future role of the robot - is assumed. As outlined before, negative feedback provided by a high reward assistant robot should lead to more negative evaluations of the robot's sociability as well as the interaction with it than negative feedback by a low reward competitor robot. These effects should be enhanced when the feedback is believed to be generated autonomously by the robot itself and diminished when the feedback is believed to be pre-programmed leaving the robot no freedom of choice over how to react.

Consequently, this hypothesis is postulated:

H7 When a social robot is expected to become an assistant, the robot's negative and believed to be autonomous feedback leads to a more negative evaluation of (a) the robot's sociability and (b) the interaction with the robot, compared to positive feedback which is believed to be pre-programmed and comes from a robot that is expected to become a competitor.

2 Method

The online study consists of an experimental 2 (positive vs. negative feedback) × 2 (autonomous vs. programmed feedback) × 2 (assistant- vs. competitor-expectation) between-subjects design. Participants were randomly assigned to one of the eight conditions. The ethics committee of the division of Computer Science and Applied Cognitive Sciences at the Faculty of Engineering of the University of Duisburg-Essen approved the study and written informed consent was obtained.

2.1 Sample

The link leading to the online study was distributed in various Facebook groups and Internet forums and sent to personal WhatsApp and Instagram contacts. Furthermore, 122 participants were recruited using the panel platform Prolific. In total, 533 people completed the online study, of which the first 131 data sets had to be excluded due to a programming error and another 8 were dismissed due to severe signs of inattention (e.g., unrealistically fast response times, conspicuous answer patterns, or failing all three manipulation checks). Regarding the manipulation checks, 331 participants passed all of them and 63 failed one or two of them (38 failed to choose the correct future role, 14 failed the question regarding the autonomy of the feedback, and 22 did not perceive the feedback the way it was intended). These 63 data sets were checked carefully regarding further signs of inattention. Since no further indications were found, we assume that these participants chose the wrong answer by mistake. Further considering the extensive wording of the answer options and the fact that there was only one item for each manipulation check, we decided to retain the respective data sets in the sample. Moreover, the main analyses were repeated with a data set from which the 63 participants who failed one or more of the manipulation checks were excluded. There were no substantial differences regarding the direction and power of the effects, which is why the full data set was used for the analyses. The comparison of effects as well as the full and reduced data sets can be viewed in the online supplementary material.

In the remaining sample of 394 subjects, 245 declared to be female, 147 to be male and two regarded themselves as diverse (see Table 1 for the sample distribution). The age range was 18 to 68 with an average of 29.66 (SD = 9.85) years. With regard to education, most of the participants reported to be in possession of university entrance-level qualifications (33.8%) or a university degree (57.1%). Most participants declared to be college students (50.8%) or employed (33.8%). There was no significant difference regarding participants' age ($p = 0.649$), sex ($p = 0.583$), education ($p = 0.732$), and employment status ($p = 0.159$) between the eight conditions.

Looking at participants' experiences with robots, 112 had no prior contact, 27 have never seen a report about real robots, and 22 have never seen a science fiction movie or series with robots. Among the remaining participants, the contact frequency with real robots was rather low ($M = 1.76$, SD = 1.06). The reception frequency of reports about real robots

Table 1 Sample distribution by experimental condition, sex and in total

Future role	Assistant				Competitor				Total
Autonomy of feedback	Autonomous		Programmed		Autonomous		Programmed		
Valence of feedback	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	
Male	21	20	19	19	15	21	19	13	147
Female	34	31	26	30	31	29	30	34	245
Diverse	0	0	1	0	1	0	0	0	2
Total	55	51	46	49	47	50	49	47	394

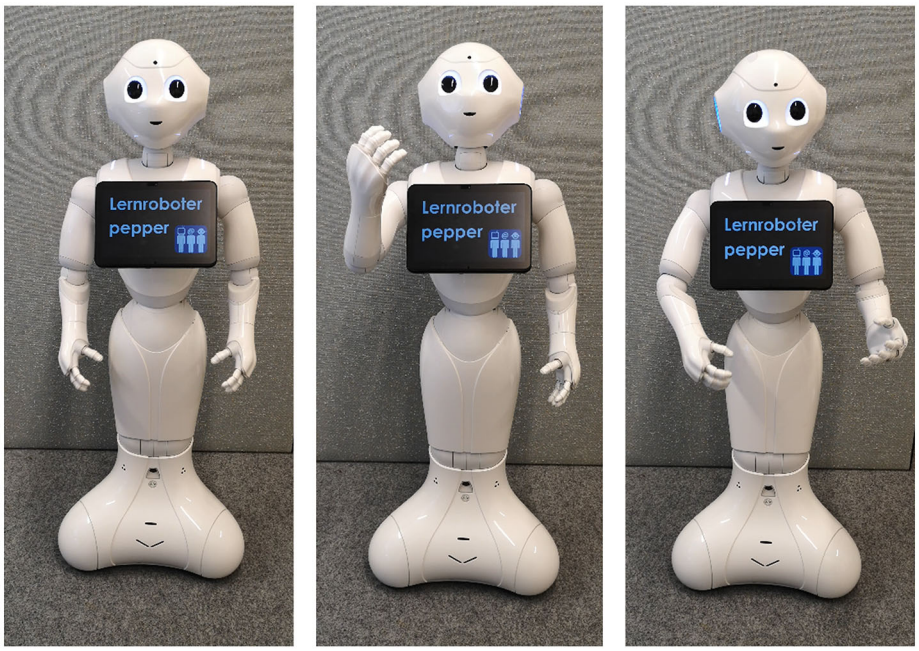

Fig. 1 Social robot Pepper as portrayed in the video

($M = 2.56$, SD $= 1.21$) and science fiction with robots ($M = 2.60$, SD $= 1.29$) was noticeably higher.

2.2 Experimental Procedure

First, people read about the aim and context of the study and their informed consent was obtained. The study started with some pretest questionnaires concerning participants' demographic and technological background. Next, a text describing the experimental manipulations regarding the robot Pepper's expected future role and its autonomy in giving feedback was presented (see Sect. 2.3 for the experimental manipulations; the vignettes are further uploaded to the online supplementary material). This was followed by Pepper introducing itself and repeating the experimental manipulations once again for reinforcement (see Fig. 1; the full interaction script and the original videos can be viewed in the online supplementary material). Although prerecorded videos were used, participants were led to believe they would be interacting live with Pepper. This was important for the believability of the manipulations, particularly in case of the robot's feedback being created autonomously by the robot. To enhance people's impression of a live interaction, a loading screen saying "establishing connection" was shown before the video started. Moreover, the robot asked how the participant was feeling and responded according to the chosen answer option (e.g., when participants answered "rather good", Pepper answered "I'm happy to hear that! I am good as well today.").

After the robot's introduction, participants watched a short video (3 min 25 s) about a scientific topic, more specifically about the soil. Afterwards, there were manipulation checks regarding the robot's believed autonomy and expected future role (see Sect. 2.4.4 for further details). Then, another video with Pepper was played to start with the quiz. The quiz consisted of Pepper asking ten questions regarding the content of the tutorial video, which were purposely chosen to be hard to answer (difficult, ambiguous, no answer options; e.g., "How big are sand grains?"). Participants entered their answers via free text input for which they had a time frame of 15 s. After

they entered their answer or the timer expired, they were transferred to the next page. Participants were told that if they give no answer, it would be rated as wrong answer. After every two answers, Pepper gave non-specific feedback to the participants regarding their general current performance. This feedback was designed to always fit roughly regardless of what the participant answered (see Sect. 2.3 for example feedback and the online supplementary material for the complete script).

After the quiz, people completed an implicit association test and questionnaires asking how they evaluate the robot Pepper's agency, sociability, and competence as well as the interaction with it in general (see Sect. 2.4 for the exact measurements). Negative expectancies regarding and previous contact with robots were also assessed. After the debriefing, all participants were given the opportunity to enter their email address to win a gift certificate (1×50 € and 10×15 €). 115 students from the university where this study was conducted were able to additionally receive course credits for their study program. The 122 participants who were recruited via the panel platform Prolific received £ 3.15 as compensation. On average, people took about 28 min to participate in the online study, of which people interacted about 9 min with Pepper, spent 13 min on the questionnaires and an additional 5 min on the automatic activation of attitudes task.

2.3 Experimental Manipulations

The exact wording of the vignettes as well as the robot's interaction script can be viewed in the online supplementary material.

2.3.1 Expected Future Role

The social robot Pepper was either described as a competitor (Pepper's skills will be superior to human skills and it will take over many tasks which are currently executed by humans, because it will be able to do them in a more efficient, reliable, and safe way) or as an assistant (Pepper will be very helpful and assist humans with many exhausting and onerous tasks, to make them easier and more pleasant to do).

2.3.2 Autonomy of Feedback

Participants were told that Pepper was trained to be a tutor and able to access and analyze their answers in order to give appropriate feedback. This feedback was supposedly either autonomous or pre-programmed. The autonomous feedback was described as the robot creating feedback by itself. Here it was emphasized that the robot decides freely what to say. In case of the pre-programmed feedback, participants were told that a human programmer previously specified which feedback is presented and when. It was further emphasized that the robot has no choice about what it will say.

2.3.3 Valence of Feedback

During the quiz, Pepper gave consistently either very positive feedback (e.g., "I am impressed by how well you are doing. You have paid close attention! That is great.") or very negative feedback (e.g., "I am impressed by how bad you are at this. Have you not paid attention at all? This is unbelievable!") after every two questions (five times in total).

2.4 Measurements

2.4.1 Participant's Background

As part of the pre-questionnaires, participants' age, sex, educational level, and current employment or training status were assessed. With regard to technological background, participants' locus of control when using technology [57] and their technical affinity [58] were measured. These technological background variables were assessed as potential further influences, but no significant effects were found. Therefore, they are not further considered in the analyses.

2.4.2 Evaluation of the Robot Pepper

All adapted and self-constructed items can be viewed in the online supplementary material. The robot's perceived *agency* was assessed with an adapted version of the Sense of Agency Scale ([59]; 11 items; e.g., "Pepper is in full control of what it does."; 1 = "strongly disagree" to 5 = "strongly agree"; $M = 2.17$, SD $= 0.75$; $\alpha = 0.86$). To measure *responsibility for feedback*, four self-constructed items asked whether the robot Pepper or the person(s) who programmed the robot are rather to be held responsible for the feedback the robot provided during the interaction (e.g., "The robot Pepper is responsible for the content of the feedback."; 1 = "strongly disagree" to 5 = "strongly agree"; $M = 1.73$, SD $= 0.81$; $\alpha = 0.80$).

In order to assess how people evaluate the robot Pepper's *sociability* after interacting with it online, an adapted version of the social attractiveness subscale of the Interpersonal Attractiveness Scale ([60]; 1 = "strongly disagree" to 5 = "strongly agree"; 5 items; e.g., "I think the robot Pepper could be a friend of mine."; $M = 2.42$, SD $= 0.97$; $\alpha = 0.79$) was used. Out of a pool of adjectives coming from several person and robot evaluation scales [49, 61–66], frequently used adjectives were identified. From these adjectives, 15 were chosen and adapted which are suitable to describe the robot's perceived sociability (e.g., "cold – warm"; $M = 3.07$, SD $= 1.03$; $\alpha = 0.96$) on a 5-point semantical differential.

The robot's perceived *competence* was measured using the equally adapted task attractiveness subscale of the Interpersonal Attractiveness Scale ([60]; 5 items; e.g., "The robot Pepper would be a poor problem solver."; $M = 3.18$, SD $= 1.03$; $\alpha = 0.89$). Furthermore, 9 adapted items from the adjective pool mentioned before were used which are fitting to measure the robot's perceived competence (e.g., "incapable – capable"; 5-point semantical differential; $M = 2.98$, SD $= 0.89$; $\alpha = 0.92$). The theoretical constructs sociability and competence were verified via factor analysis. A principal component analysis and varimax rotations of the factor loading matrix were used. All items meet the minimum criteria of having a primary factor loading of 0.4 or above and no cross-loading of 0.3 or above (further details in the online supplementary materials).

2.4.3 Evaluation of the Interaction with Robot Pepper

For a general *evaluation* of the interaction with the robot Pepper, an adapted version of the Evaluation subscale ([67]; 4 items; e.g., "I was enjoying the interaction with the robot Pepper."; 1 = "strongly disagree" to 5 = "strongly agree"; $M = 2.80$, $SD = 1.20$; $\alpha = 0.91$) was used.

2.5 Further Assessments

Manipulations were checked by asking how the robot's future role (rough answers options: assistant, replacement, information guide, or no information on future role) and how the generation of the robot's feedback (rough answer options: autonomous, pre-programmed, randomized, or I would have to guess) were described. After the interaction, it was asked how the robot's feedback was perceived (rough answer options: rather impolite, rather polite, or neutral).

Moreover the following aspects were assessed within the study but not needed for the analyses of this paper: human-likeness (self-constructed), psychological safety [68, 69], helpfulness of feedback [8], positive and negative affect [70], expectedness [67] and appropriateness of the robot's behavior (self-constructed), contact intentions [71], previous experiences with robots (based on [13]) and negative expectancies regarding robots [13].

2.5.1 Automatic Activation of Attitudes

A priming procedure was used to automatically activate attitudes in order to measure people's implicit attitudes towards robots [72]. Evaluations based on self-report may not always reflect people's underlying attitudes [72–74], which is why this procedure was used as an additional measurement. During this task, which was labeled as memory task for the participants, twelve different objects were shown for one second to the participants which worked as picture primes (cf. [74]). These objects included three robots (Pepper, Nao, and Aibo), three objects expected to produce positive evaluations (dog, butterfly, and guitar), three objects expected to produce negative evaluations (cockroach, spider, and skull) and three objects expected to produce neutral evaluations (fork, letter, and broom). A pool of positive, negative, and neutral picture primes were retrieved from Giner-Sorolla et al. [74]. They were evaluated in a pilot test (N = 15), based on which we chose the final three objects for each category. After each picture, participants were instructed to categorize 20 different adjectives (e.g., "appealing" or "repulsive"; [72, 74]) either as positive or negative by pressing a designated key on their computer keyboard ("D" for positive and "K" for negative). The adjective remained on the screen until the participant pressed one of the two keys. Participants were instructed to memorize the shown object and to respond quickly to the adjectives, but to avoid making too many mistakes. For this priming task, participants' responses and their reaction times were recorded and compared between the different picture primes. Previous research has shown that in a priming paradigm, pictures of objects which are evaluated as extremely positive speed the evaluation of same-valence targets, i.e., positively valenced adjectives, when displayed for a very short time beforehand [74, 75]. The same was shown to be true for pictures of objects which are evaluated as extremely negative and negatively valenced adjectives. The aim for this study was to assess whether pictures of robots likewise may lead to a quickened categorization of positively or negatively valenced adjectives which would suggest an automatic activation of either positive or negative attitudes towards robots.

3 Results

An overview of the descriptive values of the main influencing and dependent variables are presented in Table 2.

3.1 Autonomous vs. Pre-Programmed Feedback

To test the first hypothesis (H1), which postulates that feedback which is believed to be generated autonomously by the robot leads to an enhanced perception of (a) the robot's agency, (b) the robot being responsible for the feedback, and (c) the robot's competence compared to pre-programmed feedback, a multivariate analysis of variance (MANOVA) was calculated.

Using Pillai's trace, there was a significant main effect of the believed autonomy of the social robot's feedback on its perceived agency, responsibility for the feedback, and competence, $V = 0.14$, $F(4389) = 15.48$, $p < 0.001$. Separate univariate ANOVAs on the different outcome variables revealed a significant effect of autonomous feedback

Table 2 Descriptive statistics of the main dependent variables by feedback autonomy, feedback valence, and expected future role

	Feedback autonomy				Feedback valence				Expected future role			
	Autonomous (N = 203)		Programmed (N = 191)		Negative (N = 197)		Positive (N = 197)		Competitor (N = 193)		Assistant (N = 201)	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Agency	2.43	0.76	1.90	0.63	2.14	0.75	2.20	0.75	2.16	0.72	2.18	0.78
Responsibility	1.97	0.87	1.48	0.67	1.72	0.79	1.74	0.84	1.74	0.82	1.72	0.81
Task attract	3.21	1.03	3.15	1.02	2.82	1.00	3.55	0.92	3.21	1.02	3.15	1.04
Competence	3.07	0.89	2.88	0.88	2.80	0.91	3.16	0.83	3.01	0.82	2.95	0.95
Social attract	2.48	1.00	2.35	0.94	2.30	0.93	2.54	1.00	2.43	0.94	2.40	1.00
Sociability	3.10	0.99	3.05	1.07	2.27	0.73	3.88	0.53	3.06	1.02	3.09	1.04
Evaluation	2.77	1.19	2.82	1.22	2.14	1.01	3.46	1.00	2.78	1.17	2.81	1.24

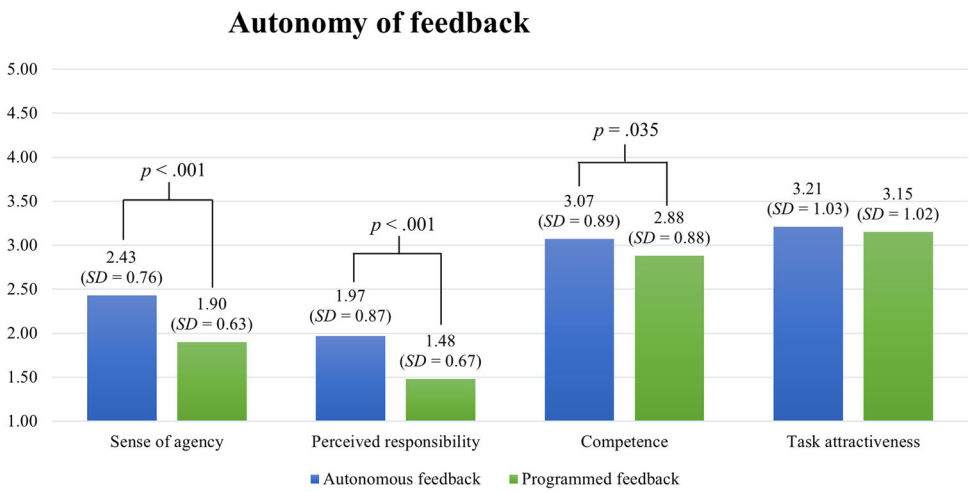

Fig. 2 Values for sense of agency, perceived responsibility, competence, and task attractiveness for autonomous and programmed feedback respectively

on the robot Pepper's perceived agency, $F(1392) = 56.60$, $p < 0.001$, $\eta_p^2 = 0.13$, supporting H1a. There was also a significant effect on its perceived responsibility for the feedback content, $F(1392) = 38.79$, $p < 0.001$, $\eta_p^2 = 0.09$, supporting H1b. Furthermore, there was a significant effect of the robot's believed feedback autonomy on competence, $F(1392) = 4.46$, $p = 0.035$, $\eta_p^2 = 0.01$, but not on task attractiveness, $F(1392) = 0.36$, $p = 0.549$, which partly supports H1c. Summing up, a robot believed to be creating feedback autonomously by itself led people to evaluate the robot's agency, its responsibility for the feedback content, as well as its competence higher than a robot which only provided strictly pre-programmed feedback (see Table 2; Fig. 2).

To test H2, which postulates a mediation effect of the robot's believed feedback autonomy via the robot's perceived agency on (a) its perceived sociability and (b) the evaluation of the interaction, three mediation analyses were calculated (number of bootstrapping samples = 5000; see Fig. 3). The criterion of the first mediation analysis was social attractiveness, for the second it was sociability, and for the third it was the evaluation of the interaction. There were significant indirect effects of the robot's believed feedback autonomy via its perceived agency respectively on social attractiveness, $b = 0.25$, 95% CI [0.16, 0.36], on sociability, $b = 0.09$, 95% CI [0.01, 0.18], and on the interaction evaluation $b = 0.22$, 95% CI [0.12, 0.34]. Consequently, H2a and H2b are supported. Direct effects of the believed feedback autonomy on social attractiveness, $b = -0.11$, $p = .253$, and on sociability, $b = -0.05$, $p = 0.658$, were not significant. However, there was a significant negative direct effect of the believed feedback autonomy on the interaction evaluation, $b = -0.27$, $p = .033$.

To explore these results further, the mediation analyses were conducted again but this time the data set was split by participants who received negative and participants who received positive feedback. For negative feedback, there were significant indirect effects of the believed feedback autonomy on social attractiveness, $b = 0.16$, 95% CI [0.06, 0.29], and on the evaluation of the interaction, $b = 0.13$, 95% CI [0.03, 0.27]. However, there was no significant indirect effect of the believed feedback autonomy on sociability, $b = 0.06$, 95% CI [0.01, 0.15] and no significant direct effects on all three criterion variables (social attractiveness: $b = 0.07$, $p = .602$; sociability: $b = 0.04$, $p = .719$; interaction evaluation: $b = -0.04$, $p = .801$). Only looking at participants who received positive feedback, there were significant indirect effects of the believed feedback autonomy on social attractiveness, $b =$

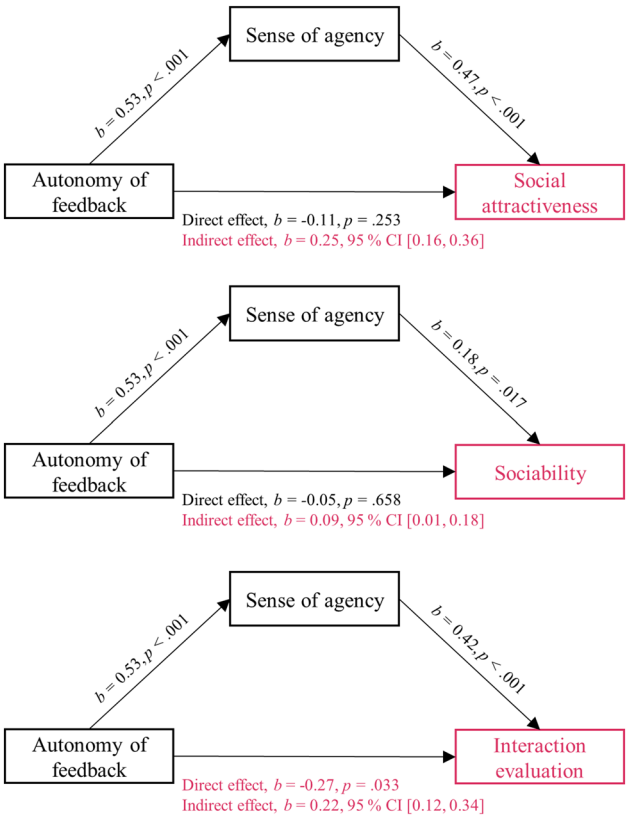

Fig. 3 Mediation models: Believed autonomy of feedback, sense of agency, and social attractiveness, sociability, or interaction evaluation respectively

0.37, 9% CI [0.23, 0.54], sociability, $b = 0.08$, 9% CI [0.01, 0.16], and the interaction evaluation, $b = 0.29$, 9% CI [0.15, 0.46]. There were significant direct effects as well on social attractiveness, $b = -0.32$, $p = .026$, and the interaction evaluation, $b = -0.47$, $p = .002$, but not on sociability, $b = -0.07$, $p = .407$. Interestingly, when people received positive feedback, direct effects of the believed feedback autonomy on the evaluation of the robot's social attractiveness and the interaction with it were negative while indirect effects were positive. Thus, a robot believed to be giving autonomous positive feedback directly leads to a more negative evaluation of the robot's social attractiveness and the general interaction with it. However, positive feedback believed to be autonomous also leads to a heightened perceived agency of the robot, which then in turn leads to more positive evaluations of the robot's sociability, social attractiveness, and the interaction with it.

3.2 Feedback Valence, Believed Feedback Autonomy, and the Robot's Expected Future Role

To test the remaining five hypotheses (H3−H7), which assume a main or interaction effect of the three manipulations (feedback valence, expected future role, and believed feedback autonomy) on the evaluation of the robot's perceived sociability and the interaction with it, another MANOVA was calculated.

The third hypothesis (H3) assumes an effect of the valence of the robot's feedback on (a) its perceived sociability and (b) the evaluation of the interaction with the robot. Using Pillai's trace, a significant effect of the feedback valence on the evaluation of the robot's sociability and the interaction with it was revealed, $V = 0.66$, $F(3, 384) = 248.61$, $p < .001$. Separate univariate ANOVAs on the outcome variables revealed a significant effect on social attractiveness, $F(1, 386) = 6.67$, $p = .010$, $\eta_p^2 = 0.02$, sociability, $F(1, 386) = 633.75$, $p < .001$, $\eta_p^2 = 0.62$, as well as interaction evaluation, $F(1, 386) = 169.32$, $p < .001$, $\eta_p^2 = 0.31$. Positive feedback led to a better evaluation of the robot's social attractiveness and sociability as well as of the interaction with the robot (see Table 2; Fig. 4). Thus, H3 is fully supported.

Hypothesis H4 further takes into account a combined influence of the feedback valence and whether it is believed to be generated autonomously or pre-programmed on (a) the robot's perceived sociability and (b) the interaction evaluation. Using Pillai's trace, there was no significant effect of the combined effect of the feedback valence and believed autonomy on the evaluation of the robot's sociability and the interaction with it, $V = 0.01$, $F(3, 384) = 0.62$, $p = .602$. Based on these results, H4 needs to be fully rejected.

The next hypothesis (H5) looks at the effect of the robot's expected future role on (a) its perceived sociability and (b) the evaluation of the interaction. Using Pillai's trace, there was no significant effect on the evaluation of the robot's sociability and the interaction with it, $V = 0.00$, $F(3, 384) = 0.29$, $p = .834$. Consequently, there is no support for H5 in the data.

Next, hypothesis H6 postulates a combined effect of the feedback valence and the robot's expected future role on (a) the robot's perceived sociability and (b) the interaction evaluation. According to Pillai's trace, again no significant effect was detected on the evaluation of the robot's sociability and the interaction with it, $V = 0.01$, $F(3, 384) = 0.95$, $p = .419$. According to these results, H6 needs to be rejected as well.

The last hypothesis (H7) considers the combined effect of all manipulation factors, i.e., the believed autonomy of the robot's feedback, the valence of the feedback, and the robot's expected future role on (a) the robot's perceived sociability and (b) the interaction evaluation. Using Pillai's trace, there was a significant effect on the evaluation of the robot's sociability and the interaction with it, $V = 0.03$, $F(3, 384) = 4.20$, $p = .006$. However, separate univariate ANOVAs only revealed a significant effect of the three-way interaction on social attractiveness, $F(1, 386) = 4.97$, $p = .026$, $\eta_p^2 = 0.01$, but not on sociability, $F(1, 386) = 1.43$, $p = .233$, and not on the interaction evaluation, $F(1, 386) = 0.12$, $p = .726$.

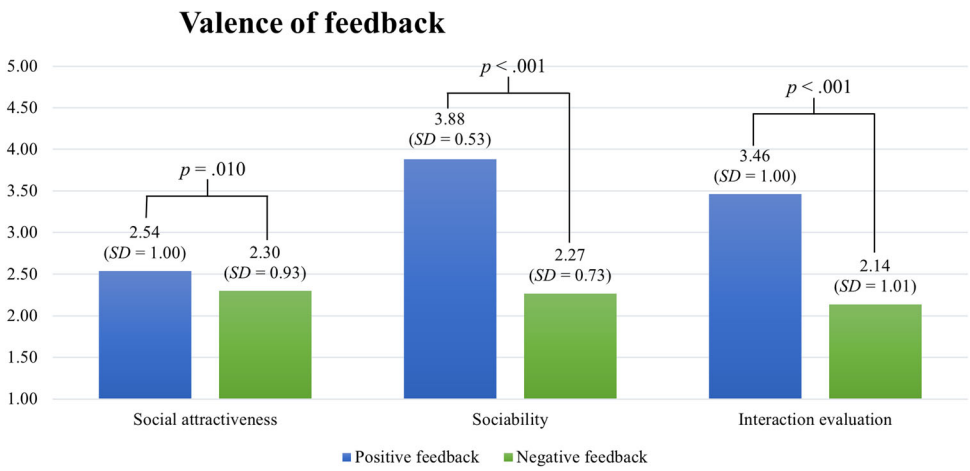

Fig. 4 Values for social attractiveness, sociability, and interaction evaluation for positive and negative feedback

A follow-up simple effect analysis for social attractiveness revealed that when the feedback was programmed and the robot was framed to become an assistant, there was a significant difference between negative and positive feedback, $p < .001$ (see Fig. 5): The robot's social attractiveness was evaluated higher when the feedback was positive ($M = 2.63$, SD = 0.96) compared to when it was negative ($M = 1.91$, SD = 0.78). Therefore, H7a is partly supported (with regard to social attractiveness). H7b is rejected.

3.2.1 Automatic Activation of Attitudes

To receive some further results, a mixed design ANOVA was calculated to analyze the results of the implicit measure of automatic activation of attitudes. Since the automatic activation of attitudes task could only be completed via computer, the task was not presented to individuals participating via smartphone. As a result, 335 data sets were available for these analyses. Error responses (wrong classification of adjectives; 8.86% of all responses) as well as responses lower than 300 ms (2.45% of all correct responses) or greater than 2.500 ms (1.36% of all correct responses) were excluded from analyses following the procedure by Giner-Sorolla et al. [74].

Mauchly's test indicated that the assumption of sphericity was violated for the interaction effect of the prime picture (robot vs. positive vs. negative vs. neutral prime picture) and the target adjective valence, $X^2(5) = 27.11$, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.94$ for the interaction effect of prime picture and target adjective valence).

There was a significant interaction effect between the target adjective valence and the type of prime picture, $F(2.83, 922.64) = 3.62$, $p = .014$, $\eta_p^2 = 0.01$, indicating that participants' response times after different types of prime pictures differed between positive and negative target adjectives. Contrasts were performed comparing robot, positive, and negative prime pictures to neutral prime pictures across negative and positive target adjectives. These contrasts revealed a significant interaction when comparing positive to neutral prime pictures, $F(1, 326) = 12.48$, $p < .001$, $\eta_p^2 = 0.04$. The interaction graph and estimated marginal means show that after seeing a positive prime picture, participants responded faster to positive target adjectives ($EMM = 795.63$, $SE = 7.27$) compared to a neutral prime picture ($EMM = 812.74$, $SE = 8.75$). Likewise, a positive prime picture leads to greater response times for negative target adjectives ($EMM = 818.82$, $SE = 7.45$) compared to the effects of a neutral prime ($EMM = 811.42$, $SE = 7.57$). Contrasts revealed no significant interaction when comparing negative ($F(1326) = 1.71$, $p = .192$) as well as robot ($F(1326) = 1.37$, $p = .243$) prime pictures to neutral prime pictures.

Further, there was no significant interaction effect of the target adjective valence, the type of prime picture, and the valence of the robot's feedback, $F(2.83, 922.64) = 0.92$, $p = .427$. This indicates that there was no significant difference regarding the robot's positive or negative feedback in participants' response times for the different types of prime pictures, neither for positive nor for negative target adjectives.

4 Discussion

Since successful interactions with humans are pivotal for the further development of social robots, frustrations and misunderstandings should be avoided. To reach this goal, further insights are necessary to understand the mechanisms that take place in human-robot interactions. Against this background, the aim of this study was to explore which circumstances may affect whether and to what extent responsibility for a social robot's actions is attributed to the robot and how this affects the evaluation of the robot and the interaction with it. For this purpose, participants were told that they will either

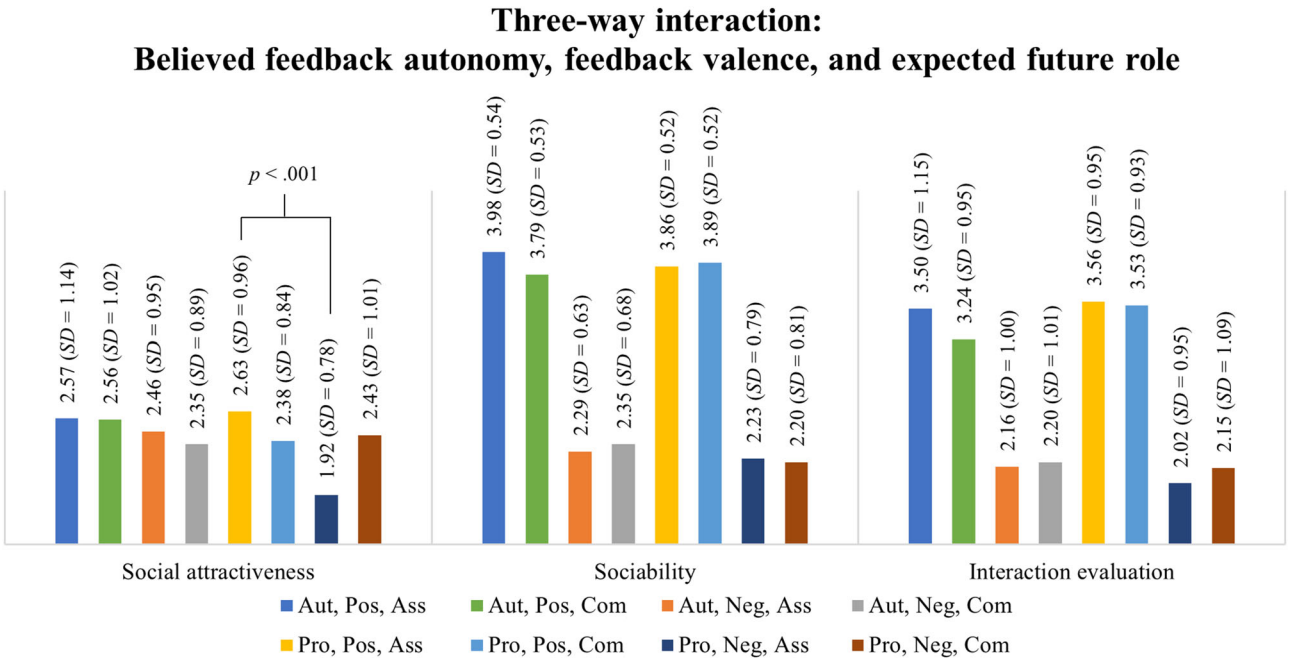

Fig. 5 Three-way interaction of the believed feedback autonomy, the feedback valence, and the robot's expected future role (Aut = autonomous, Pro = programmed, Pos = Positive, Neg = Negative, Ass = Assistant, Com = Competitor)

receive feedback which has been pre-programmed or which is generated autonomously by the robot itself. In addition to examining how the presence of an external justification for the robot's behavior affects people's judgements, it was further considered whether the valence of the feedback (positive vs. negative) and the robot's expected future role (assistant vs. competitor) have an influence.

4.1 Autonomous vs. Pre-Programmed Feedback

Autonomy is connected to internal locus of causality [21, 22], for which a certain amount of agency, i.e. the capacity to act independently and free of choice, is a prerequisite [23]. Objectively speaking, robots are neither completely autonomous, since their actions usually depend on others such as users or programmers, nor in possession of full agency, since they lack consciousness, intention, and a free will [27]. However, interactive devices are often treated as if they were living beings [28] and thus it is more important to ask how autonomous a robot is perceived or believed to be than whether it can actually be defined as autonomous.

Confirming our assumptions, results show that when the robot is believed to give feedback autonomously, more agency as well as responsibility for the content of the feedback is attributed to it, compared to when the feedback is believed to be determined by programmers beforehand. According to the effect sizes, these were the strongest relationships found in the analyses. Likewise, in case of feedback that is believed to be generated autonomously, the robot is evaluated as more competent (however, not with regard to task attractiveness and also with a rather small effect size). Since autonomy is perceived to stem from agency[22] and own intentions [29], this may explain why people ascribe more agency, responsibility, as well as competence to a robot which they believe creates feedback autonomously. This is also in line with previous research showing that the more autonomous a robot is perceived, the more responsibility is attributed to it [16, 30]. In accordance with attribution theory, results indicate that people used the information available to them to decide who is responsible for the content of the feedback [2, 3].

Results further show a full mediation indicating that feedback believed to be autonomous compared to pre-programmed increases the robot's perceived agency which then enhances how sociable the robot and how positive the interaction with it is evaluated. However, how the feedback is generated has no significant direct effect on the evaluation of the robot's sociability and even a significant negative direct effect on the evaluation of the interaction. Looking at the effects for negative and positive feedback respectively, positive feedback believed to be generated autonomously directly leads to a more negative evaluation of the robot's social attractiveness and the interaction with it, despite the positive indirect effects. Many participants reported that the positive feedback did not match their own impression of their performance which caused them to perceive the feedback as unfitting and thus neither credible nor trustworthy. In previous research, people were found to prefer accurate

feedback about themselves over positive feedback [76]. They also seek consistent information particularly when they are certain about themselves and feedback challenges these self-perceptions [77]. Since participants may have expected more accuracy in case of the autonomously generated compared to the strictly pre-programmed feedback, this may explain why for positive feedback, feedback believed to be autonomous led to worsened evaluation outcomes. The sociability evaluation was probably not affected by this since the items here ask about general characteristics of the robot, for example whether it is rather rude or kind, impolite or polite, which was likely directly affected by the robot's positive or negative feedback. This is also reflected by the effect sizes which show substantially greater effects of the feedback valence on the sociability and interaction evaluation than on the social attractiveness evaluation. The social attractiveness items ask whether the participants would like to be friends with Pepper, which they probably rather decline in case of perceived insincerity due to inadequate feedback. Furthermore, the positive, pre-programmed feedback could have been perceived to rather come from a person (since it was programmed by a person) which due to people's pronounced need to belong [53] might have been evaluated better than when it is believed to come directly from the robot.

With negative feedback, there was no direct effect of the believed autonomy of the feedback on any evaluation measures. This could be due to the intensity of feelings elicited by the negative feedback, which may push the influence of any other justifications, such as whether the feedback is generated autonomously or pre-programmed, aside. Previous research already points to the central influence of behavior in interactions with artificial entities [15, 78].

4.2 Valence of Feedback

When looking at the effects of the feedback valence alone, negative feedback clearly causes people to evaluate the robot's sociability as well as the general interaction with it more negatively compared to positive feedback. This extends previous findings that negative feedback by an artificial entity leads to a detrimental evaluation of the entity [7, 8]. On the one hand, this may be explained by people's need to protect their self-esteem, which was shown to result in a de-evaluation of others [37]. Furthermore, several studies deliver evidence for reciprocal, sometimes even retaliating behavior in human-computer interactions [45–50]. Thus, a simple explanation might be that people just evaluated the robot poorly because they were evaluated poorly by the robot. Furthermore, due to their field of application and their human-like characteristics, interactive devices are expected to be polite and friendly [38]. Considering this, negative feedback should also violate people's expectations in a clearly negative way, which usually results in detrimental communication and relationship outcomes [39]. In our case, the negative expectancy violation may have caused people to evaluate the robot's sociability and the general interaction with it poorly.

4.3 Believed Autonomy Interacting with Feedback Valence

People pervasively tend to overestimate the influence of internal factors, while underestimating external factors, which is described as fundamental attribution error [2, 18, 19]. This study's predominant focus was to explore whether an external justification for the robot's behavior affects how the robot is evaluated after giving negative or positive feedback. Since direct effects of the valence of the feedback on evaluations of the robot's sociability as well as the interaction with it were found which appear not to be affected by the external justification that the feedback was not created by the robot but programmed in advance, this suggests the occurrence of a fundamental attribution error [2]. Results indicate that when evaluating the robot, it was not taken into account whether the robot created the feedback itself or whether the feedback was programmed in advance by someone else. Negative feedback led to a poor evaluation regardless of the information whether the robot was the creator or the transmitter of the feedback. This points to an underestimation of the situational constraints under which the robot was acting and an overestimation of the robot's (not existent) intentions behind the feedback, i.e., the fundamental attribution error [2].

4.4 Robots' Expected Future Roles

In addition to the valence and the believed autonomy of the robot's feedback, this study further examined whether framing the robot in a positive or negative light might influence the evaluation of the robot's sociability and the interaction with it as well. For this purpose, the two most prominent expectations for social robots' future roles were emphasized, one pleasant and desired, the other unpleasant and worried about [9]. Consequently, the robot Pepper was either described as a helpful assistant or as a threatening competitor [11–13].

Although previous research showed that describing the robot either as an assistant or a competitor affects people's evaluation of the robot's sociability and the interaction with it [15], no main or interaction effects of the robot's expected future role were found in this study. Looking at the effect sizes of the feedback valence on sociability and interaction evaluations, an explanation could be that the effect of the robot's behavior was so strong that the previous description of the robot's future role was superposed and did not play a decisive role when evaluating the robot. People experienced the robot's behavior themselves while the expected role was only described to them. Moreover, which role the

robot aspires may be perceived as distant future and thus not as present as the behavior displayed in that moment [15]. As Rickenberg and Reeves [78] point out, an agent's behavior during an interaction plays a pivotal role for a subsequent evaluation of the agent.

4.5 Automatic Activation of Attitudes

Implicit measures showed that participants responded faster to positive and slower to negative target adjectives after seeing a positive compared to neutral prime pictures, which confirms previous results [72, 74]. However, no differences in the response times for classifying negative and positive target adjectives were found when comparing negative as well as robot prime pictures to neutral prime pictures. There was also no significant interaction effect of the target adjective valence, the type of prime picture, and the valence of the robot's feedback. Thus, whether the robot gives positive or negative feedback appears not to affect how people implicitly evaluate the robot. However, it is interesting that positive prime pictures led to the intended difference in response times, while robot prime pictures did not differ significantly from neutral pictures. An explanation could be that the robots we showed (Pepper, Nao, and Aibo) are generally evaluated neutrally since they are clearly identifiable as robots which people state to prefer over too realistic human-like robots [11, 13]. Furthermore, the one-time online interaction with the robot Pepper probably did not change people's attitude toward robots in general even if they received harsh, negative feedback by one of their specimens.

4.6 Limitations and Future Research

In this study, self-reported variables were extended by implicit measures. However, additional behavioral measurements may bring further insights, for example regarding verbal and non-verbal expressions, reciprocity, and retaliation behaviors. Since the questionnaires were presented in the same sequence to all participants, order effects may have occurred. Moreover, some of the obtained effect sizes were strikingly low which may indicate a limited explanatory power.

Another aspect which needs to be mentioned is the failure of manipulation checks, which may mean that not all participants remembered the robot's future role and how its feedback is generated correctly as well as that the valence of the feedback was not always perceived as intended. However, manipulations were reinforced in written form as well as orally by the robot itself, which gives us the confidence that all participants received the information, even though some failed to recall it correctly later. In future studies, the manipulations may also be reinforced by behavioral means.

Several participants further reported that they perceived the positive feedback to be incredible and inaccurate since it did not match how they evaluated their own performance. This may have affected the evaluation of the robot and the interaction with it as well. Therefore, in future studies the perceived credibility and accuracy of the feedback should be assessed. For this study, specific feedback in accordance with the participant's actual performance would not have allowed Pepper to show either consistently positive or negative behavior, which would have counteracted the study's research aim. However, we recommend that future studies control for the feedback to be similarly accurate across conditions. For this purpose, a different task (e.g., an estimation task) may be more suitable.

Considering different contexts and environments as well as long-term effects would further extend our insights. Nevertheless, we would like to emphasize that this study used well thought out videos of a real social robot with whom people thought to be interacting with online. This was done to receive realistic reactions.

5 Conclusion

With the rapid advancement of social robots, multi-facetted and socially situated interactions with humans in their everyday environments become increasingly common [1]. For a successful design of these interactions while avoiding misunderstandings, errors, and subsequent frustrations, this study considers insights of attribution theory [2, 3] to examine whether the perceived autonomy of a social robot's actions causes people to attribute agency and responsibility to the robot and how this together with the valence of the behavior and the robot's expected future role affects how the robot and the interaction with it are evaluated.

Summing up, when the feedback was believed to be created autonomously by a social robot, more agency, responsibility, as well as competence were attributed to the robot compared to when the feedback was believed to be pre-programmed. Moreover, the more agency a robot is ascribed to, the better the evaluation of its sociability and the interaction with it. However, only the valence of the feedback affected the evaluation of the robot's sociability and the interaction with it directly. Since a robot giving negative feedback was also evaluated more negatively than a robot giving positive feedback when the feedback was programmed by some programmers beforehand, this points to the occurrence of the fundamental attribution error [2]. The external justification of the pre-programmed feedback should have led to an external attribution of the negative, unpleasant behavior [3]. However, looking at the evaluation measures, the negative feedback appears to be attributed internally to the robot. Thus, when designing social robots, it should be kept in mind that the

believed autonomy of a robot might affect the attribution of agency, responsibility, and competence to it, however, the valence of its behavior alone decides how the robot's sociability and the interaction with it are evaluated.

Authors' contributions Conceptualization: ACH, NCK; Methodology: ACH, NCK; Formal analysis and investigation: ACH; Writing—original draft preparation: ACH; Writing—review and editing: ACH, NCK; Funding acquisition: not applicable; Resources: NCK; Supervision: NCK.

Funding Open Access funding enabled and organized by Projekt DEAL. No funds, grants, or other support was received.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethical approval and consent The ethics committee of the division of Computer Science and Applied Cognitive Sciences at the Faculty of Engineering of the University of Duisburg-Essen approved the study and written informed consent was obtained.

References

1. Young JE, Sung J, Voida A et al (2011) Evaluating human-robot interaction: Focusing on the holistic interaction experience. Int J Social Robot 3:53–67. https://doi.org/10.1007/s12369-010-0081-8
2. Ross L (1977) The intuitive psychologist and his shortcomings: Distortions in the attribution process. Adv Exp Soc Psychol 10:173–220. https://doi.org/10.1016/S0065-2601(08)60357-3
3. Kelley HH (1973) The processes of causal attribution. Am Psychol 28:107–128. https://doi.org/10.1037/h0034225
4. Nass CI, Moon Y, Carney P (1999) Are people polite to computers? Responses to computer-based interviewing systems. J Appl Soc Psychol 29:1093–1109. https://doi.org/10.1111/j.1559-1816.1999.tb00142.x
5. Fogg BJ, Nass CI (1997) Silicon sycophants: the effects of computers that flatter. Int J Hum Comput Stud 46:551–561. https://doi.org/10.1006/ijhc.1996.0104
6. Johnson D, Gardner J, Wiles J (2004) Experience as a moderator of the media equation: The impact of flattery and praise. Int J Hum Comput Stud 61:237–258. https://doi.org/10.1016/j.ijhcs.2003.12.008
7. Carolus A, Muench R, Schmidt C et al (2019) Impertinent mobiles - Effects of politeness and impoliteness in human-smartphone interaction. Comput Hum Behav 93:290–300. https://doi.org/10.1016/j.chb.2018.12.030
8. Krämer NC, Leiße L-M, Hollingshead A et al (2017) Evaluated by a machine: Effects of negative feedback by a computer or human boss. In: Beskow J, Peters C, Castellano G (eds) Intelligent Virtual Agents: Proceedings of the 17th International Conference on Intelligent Virtual Agents - IVA '17. Springer, Cham, Switzerland, pp 235–238
9. Bruckenberger U, Weiss A, Mirnig N et al (2013) The good, the bad, the weird: Audience evaluation of a "real" robot in relation to science fiction and mass media. In: Jamshidi M (ed) Advance Trends in Soft Computing: Proceedings of the World Conference on Soft Computing - WCSC '13, vol 8239. Springer, Cham, Switzerland, pp 301–310
10. Khan Z (1998) Attitudes towards intelligent service robots. Royal Institute of Technology, Stockholm, Sweden
11. Ray C, Mondada F, Siegwart R (2008) What do people expect from robots? In: Proceedings of the 21st IEEE/RSJ International Conference on Intelligent Robots and Systems - IROS '08. IEEE, Piscataway, NJ, pp 3816–3821
12. Weiss A, Igelsböck J, Wurhofer D et al (2011) Looking forward to a "robotic society"? Int J Social Robot 3:111–123. https://doi.org/10.1007/s12369-010-0076-5
13. Horstmann AC, Krämer NC (2019) Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. Front Psychol 10:939. https://doi.org/10.3389/fpsyg.2019.00939
14. Oestreicher L, Eklundh K (2006) User expectations on human-robot co-operation. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication - RO-MAN '06. IEEE, Piscataway, NJ, pp 91–96
15. Horstmann AC, Krämer NC (2020) Expectations vs. actual behavior of a social robot: An experimental investigation of the effects of a social robot's interaction skill level and its expected future role on people's evaluations. PLoS ONE 15:e0238133. https://doi.org/10.1371/journal.pone.0238133
16. Kim T, Hinds PJ (2006) Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication - RO-MAN '06. IEEE, Piscataway, NJ, pp 80–85
17. Hewstone M, Jaspars J (1987) Covariation and causal attribution: A Logical Model of the intuitive analysis of variance. J Personal Soc Psychol 53:663–672. https://doi.org/10.1037/0022-3514.53.4.663
18. Jones EE, Nisbett RE (1972) The actor and the observer: Divergent perceptions of the causes of behavior. In: In: Jones EE, Kanouse DE, Kelley HH et al (eds) Attribution: Perceiving the causes of behavior. General Learning Press, Morristown, NJ, pp 79–95
19. Nisbett RE, Ross L (1983) Human inference: Strategies and shortcomings of social judgment. Prentice Hall, Englewood Cliffs, NJ
20. Jones EE, Harris VA (1967) The attribution of attitudes. J Exp Soc Psychol 3:1–24. https://doi.org/10.1016/0022-1031(67)90034-0
21. de Charms R (1983) Personal causation: The internal affective determinants of behavior. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ
22. Ryan RM, Deci EL (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 55:68–78
23. Barker C (2005) Cultural studies: Theory and practice. Sage, London, UK
24. Banks J (2019) A perceived moral agency scale: Development and validation of a metric for humans and social machines. Com-

put Hum Behav 90:363–371. https://doi.org/10.1016/j.chb.2018.08.028
25. Brown LA, Walker WH (2008) Prologue: Archaeology, animism and non-human agents. J Archaeol Method Theory 15:297–299. https://doi.org/10.1007/s10816-008-9056-6
26. Allen C, Wallach W, Smit I (2006) Why machine ethics? IEEE Intell Syst 21:12–17. https://doi.org/10.1109/MIS.2006.83
27. Himma KE (2009) Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? Ethics Inf Technol 11:19–29. https://doi.org/10.1007/s10676-008-9167-5
28. Reeves B, Nass CI (1996) The media equation: How people treat computers, television, and new media like real people and places. CSLI Publications, Stanford, CA
29. Bartneck C, Forlizzi J (2004) A design-centred framework for social human-robot interaction. In: Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication - RO-MAN '04. IEEE, Piscataway, NJ, pp 591–594
30. Hinds PJ, Roberts TL, Jones H (2004) Whose job is it anyway? A study of human-robot interaction in a collaborative task. Human-Computer Interact 19:151–181. https://doi.org/10.1080/07370024.2004.9667343
31. Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. Science 315:619. https://doi.org/10.1126/science.1134475
32. Forlizzi J, DiSalvo C (2006) Service robots in the domestic environment. In: Goodrich MA, Schultz AC, Bruemmer DJ (eds) Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction - HRI '06. ACM Press, New York, NY, pp 258–265
33. Friedman B, Kahn PH, Hagman J (2003) Hardware companions? In: Cockton G, Korhonen P (eds) Proceedings of the conference on Human factors in computing systems - CHI '03. ACM Press, New York, NY, pp 273–280
34. Vallerand RJ, Reid G (1984) On the causal effects of perceived competence on intrinsic Motivation: A test of cognitive evaluation theory. J Sport Psychol 6:94–102. https://doi.org/10.1123/jsp.6.1.94
35. Kluger AN, DeNisi A (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 119:254–284. https://doi.org/10.1037/0033-2909.119.2.254
36. Krenn B, Würth S, Hergovich A (2013) The impact of feedback on goal setting and task performance. Swiss J Psychol 72:79–89. https://doi.org/10.1024/1421-0185/a000101
37. Fein S, Spencer SJ (1997) Prejudice as self-image maintenance: Affirming the self through derogating others. J Personal Soc Psychol 73:31–44. https://doi.org/10.1037/0022-3514.73.1.31
38. Sayin E, Krishna A (2019) You can't be too polite, Alexa! Implied politeness of mechanized auditory feedback and its impact on perceived performance accuracy. In: In: Bagchi R, Block L, Lee L (eds) Advances in consumer research, vol 47. Association for Consumer Research, Duluth, MN, pp 243–248
39. Burgoon JK, Hale JL (1988) Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. Communication Monographs 55:58–79. https://doi.org/10.1080/03637758809376158
40. Aula A, Surakka V (2002) Auditory emotional feedback facilitates human-computer interaction. In: Faulkner X, Finlay J, Détienne F (eds) People and Computers XVI - Memorable Yet Invisible: Proceedings of HCI 2002. Springer, London, UK, pp 337–349
41. Midden C, Ham J (2009) Using negative and positive social feedback from a robotic agent to save energy. In: Chatterjee S (ed) Proceedings of the 4th International Conference on Persuasive Technology. ACM, New York, NY, p 1
42. Ruchsow M, Grothe J, Spitzer M et al (2002) Human anterior cingulate cortex is activated by negative feedback: evidence from event-related potentials in a guessing task. Neurosci Lett 325:203–206. https://doi.org/10.1016/S0304-3940(02)00288-4
43. Aghaei Pour P, Hussain MS, AlZoubi O et al (2010) The impact of system feedback on learners' affective and physiological states. In: Aleven V, Kay J, Mostow J (eds) Proceedings of the 10th International Conference on Intelligent Tutoring Systems - ITS '10, vol 6094. Springer, Berlin, pp 264–273
44. Gouldner AW (1960) The norm of reciprocity: A preliminary statement. Am Sociol Rev 25:161. https://doi.org/10.2307/2092623
45. Fogg BJ, Nass CI (1997) How users reciprocate to computers: An experiment that demonstrates behavior change. In: In: Edwards A, Pemberton S (eds) CHI '97 extended abstracts on Human factors in computing systems looking to the future - CHI '97. ACM Press, New York, New York, USA, pp 331–332
46. Sandoval EB, Brandstetter J, Obaid M et al (2016) Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. Int J Social Robot 8:303–317. https://doi.org/10.1007/s12369-015-0323-x
47. Lorenz T, Weiss A, Hirche S (2016) Synchrony and reciprocity: Key mechanisms for social companion robots in therapy and care. Int J Social Robot 8:125–143. https://doi.org/10.1007/s12369-015-0325-8
48. Huang L, Morency L-P, Gratch J (2011) Virtual rapport 2.0. In: Hutchison D, Kanade T, Kittler J (eds) Intelligent Virtual Agents: Proceedings of the 11th International Conference on Intelligent Virtual Agents - IVA '11, vol 6895. Springer, Berlin/Heidelberg, Germany, pp 68–79
49. von der Pütten AM, Krämer NC, Gratch J et al (2010) "It doesn't matter what you are!" Explaining social effects of agents and avatars. Comput Hum Behav 26:1641–1650. https://doi.org/10.1016/j.chb.2010.06.012
50. Krämer NC, Kopp S, Becker-Asano C et al (2013) Smile and the world will smile with you - The effects of a virtual agent's smile on users' evaluation and behavior. Int J Hum Comput Stud 71:335–349. https://doi.org/10.1016/j.ijhcs.2012.09.006
51. Asher DE, Zaldivar A, Barton B et al (2012) Reciprocity and retaliation in social games with adaptive agents. IEEE Trans Auton Ment Dev 4:226–238. https://doi.org/10.1109/TAMD.2012.2202658
52. Mell J, Lucas GM, Gratch J (2018) Welcome to the real world: How agent strategy increases human willingness to deceive. In: André E, Koenig S (eds) Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems - AAMAS '18. IFAAMAS, Richland, SC, pp 1250–1257
53. Baumeister RF, Leary MR (1995) The need to belong: Desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 117:497–529
54. Asimov I (1947) Little lost robot. Street & Smith, New York, NY
55. Burgoon JK (1993) Interpersonal expectations, expectancy violations, and emotional communication. J Lang Social Psychol 12:30–48
56. Horstmann AC, Krämer NC (2020) When a Robot Violates Expectations. In: Belpaeme T, Young J, Gunes H (eds) Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction - HRI '20. ACM, New York, NY, pp 254–256
57. Beier G (1999) Kontrollüberzeugungen im Umgang mit Technik [Locus of control when using technology]. Rep Psychol 9:684–693
58. Karrer K, Glaser C, Clemens C et al (2009) Technikaffinität erfassen: Der Fragebogen TA-EG [Measuring technical affinity - the questionnaire TA-EG]. Der Mensch im Mittelpunkt Tech Syst 8:196–201
59. Tapal A, Oren E, Dar R et al (2017) The sense of agency scale: a measure of consciously perceived control over one's mind, body, and the immediate environment. Front Psychol 8:1552. https://doi.org/10.3389/fpsyg.2017.01552

60. McCroskey JC, McCain TA (1974) The measurement of interpersonal attraction. Speech Monographs 41:261–266. https://doi.org/10.1080/03637757409375845
61. Lea M, Spears R (1992) Paralanguage and social perception in computer-mediated communication. J Organ Comput 2:321–341. https://doi.org/10.1080/10919399209540190
62. Bente G, Feist A, Elder S (1996) Person perception effects of computer-simulated male and female head movement. J Nonverbal Behav 20:213–228. https://doi.org/10.1007/BF02248674
63. Fogg BJ, Tseng H (1999) The elements of computer credibility. In: Williams MG (ed) Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, New York, NY, pp 80–87
64. McCroskey JC, Young TJ (1981) Ethos and credibility: the construct and its measurement after three decades. Central States Speech J 32:24–34. https://doi.org/10.1080/10510978109368075
65. Carpinella CM, Wyman AB, Perez MA et al (2017) The Robotic Social Attributes Scale (RoSAS): Development and validation. In: Mutlu B, Tscheligi M, Weiss A (eds) Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction - HRI '17. IEEE, Piscataway, NJ, pp 254–262
66. Bartneck C, Kulić D, Croft E et al (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Social Robot 1:71–81. https://doi.org/10.1007/s12369-008-0001-3
67. Burgoon JK, Walther JB (1990) Nonverbal expectancies and the evaluative consequences of violations. Hum Commun Res 17:232–265. https://doi.org/10.1111/j.1468-2958.1990.tb00232.x
68. Baer M, Frese M (2003) Innovation is not enough: climates for initiative and psychological safety, process innovations, and firm performance. J Organ Behav 24:45–68. https://doi.org/10.1002/job.179
69. Edmondson A (1999) Psychological safety and learning behavior in work teams. Adm Sci Q 44:350–383. https://doi.org/10.2307/2666999
70. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: The PANAS scales. J Personal Soc Psychol 54:1063–1070
71. Eyssel F, Kuchenbrandt D, Bobinger S (2011) Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In: Billard A, Kahn PH, Adams JA (eds) Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction - HRI '11. ACM Press, New York, NY, pp 61–68
72. Fazio RH, Sanbonmatsu DM, Powell MC et al (1986) On the automatic activation of attitudes. J Personal Soc Psychol 50:229–238. https://doi.org/10.1037/0022-3514.50.2.229
73. Karpinski A, Hilton JL (2001) Attitudes and the Implicit Association Test. J Personal Soc Psychol 81:774–788. https://doi.org/10.1037/0022-3514.81.5.774
74. Giner-Sorolla R, García MT, Bargh JA (1999) The automatic evaluation of pictures. Soc Cogn 17:76–96. https://doi.org/10.1521/soco.1999.17.1.76
75. Hermans D, de Houwer J, Eelen P (1994) The affective priming effect: automatic activation of evaluative information in memory. Cogn Emotion 8:515–533. https://doi.org/10.1080/02699939408408957
76. Jussim L, Yen H, Aiello JR (1995) Self-consistency, self-enhancement, and accuracy in reactions to feedback. J Exp Soc Psychol 31:322–356. https://doi.org/10.1006/jesp.1995.1015
77. Pelham BW (1991) On confidence and consequence: The certainty and importance of self-knowledge. J Personal Soc Psychol 60:518–530
78. Rickenberg R, Reeves B (2000) The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In: Turner T (ed) Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, New York, NY, pp 49–56

Aike C. Horstmann received her M.Sc. in Applied Cognitive and Media Science in 2017. In 2021 she completed her Ph.D. at the Faculty of Social Psychology: Media and Communication at the University of Duisburg-Essen, Germany. With her research she focuses on human-computer interaction, i.e., interactions with social robots and virtual agents, with a special interest on expectations, attributions, and behavior. She also led various student research projects and taught courses on human-computer interaction at the University of Duisburg-Essen.

Nicole C. Krämer is Full Professor of Social Psychology: Media and Communication at the University of Duisburg-Essen, Germany. She completed her PhD in Psychology at the University of Cologne, Germany, in 2001 and received the venia legendi for psychology in 2006. Dr. Krämer´s research focuses on social psychological aspects of human-machine-interaction (especially social effects of robots and virtual agents) and computer-mediated-communication (CMC). She heads numerous projects that received third party funding. She served as Editor-in-Chief of the Journal of Media Psychology 2015–2017 and currently is Associate Editor of the Journal of Computer Mediated Communication (JCMC).