Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior

Sonja Stange
CITEC, Bielefeld University
Bielefeld, Germany
sstange@techfak.uni-bielefeld.de

Stefan Kopp
Faculty of Technology, Bielefeld University
Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

ABSTRACT

Social robots interacting with users in real-life environments will often show surprising or even undesirable behavior. In this paper we investigate whether a robot's ability to self-explain its behavior affects the users' perception and assessment of this behavior. We propose an explanation model based on humans' folk-psychological concepts and test different explanation strategies in specifically designed HRI scenarios with robot behaviors perceived as intentional, but differently surprising or desirable. All types of explanation strategies increased the understandability and desirability of the behaviors. While merely stating an action had similar effects as giving a reason for it (an intention or need), combining both in a causal explanation helped the robot to better justify its behavior and to increase its understandability and desirability to a larger extent.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; *User studies*; *Scenario-based design*; *Empirical studies in interaction design.*

KEYWORDS

Human-Robot Interaction; Behavior Explanations; Perception Study

ACM Reference Format:
Sonja Stange and Stefan Kopp. 2020. Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3319502.3374802

1 INTRODUCTION

Social robots are entering our everyday lives and have been introduced, e.g., to schools, stores or households. Thus, they come to interact with human users in a variety of different situations and over fairly long periods of time. It is highly likely – if not inevitable – that these robots often behave in ways that are not expected or even unwanted by their users (fig. 1). One approach the field has been pursuing to avoid this, is to design and implement for better interaction abilities of robots. A complementary approach, which has received growing attention recently, is to endow robots with abilities for explaining their behavior. Such explanations may aim at mitigating surprise and negative appraisal, or at establishing an understanding that enables a user to adjust the robot's behavior to personal preferences [13]. To that end, the robot's explanations must render its behavior transparent, reduce uncertainty, allow for evaluating its appropriateness under given circumstances, and eventually increase trust towards the robot [12, 15, 25].

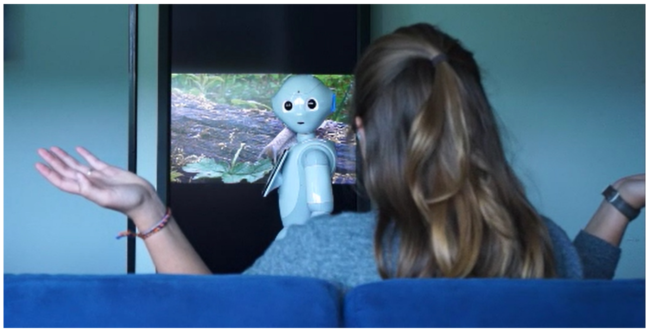

Figure 1: Example of a robot's behavior that is neither understandable nor desirable to the user, raising the need for an explanation which could be provided by the robot itself.

The present work is part of a research program to develop socially intelligent robots that can co-inhabit an environment with human users (e.g. their home). There, the robot should be able to explain its own behavior upon request, in a way that increases transparency and, eventually, enables users to tailor the robot's behavior to their individual preferences. This involves tackling two problems: First, understanding how users perceive and respond to a robot's explanation of its (possibly surprising or undesirable) behavior, and how this may change the users' current evaluation of the robot's behavior; second, developing an explanation model that is rooted in the robot's decision-making architecture to select events that have been causally relevant for its behavior, and then providing strategies to turn those into explanations that are understandable and reasonable to the user. We focus on purposeful robot behavior towards which humans tend to adopt an *intentional stance* [27] and to build explanations in terms of (human-like) *reasons* such as intentions, desires or beliefs.

In this paper, we first discuss related work in HRI as well as explanation theories from social sciences (Section 2). We propose an explanation model that offers different explanation types corresponding to a human folk-psychological concept of explanations [17]. In Section 3, we formulate hypotheses about the effects of

HRI '20, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3374802

these robot explanations and present two studies to test them: First, a preparatory study is reported (Section 3.1) that validates a set of human–robot interaction scenarios with to-be-explained behaviors, designed to be perceived as being intentional but at varying degrees of surprisingness and desirability. Then, we present results of a user study (Section 3.2) on how different explanations, when given by the robot in these interaction scenarios, justify the robot's behaviors differently well and affect the user's rating of understanding and desirability of the behaviors.

2 BACKGROUND

2.1 Related Work

The question of how to explain the decisions or actions taken by autonomous AI systems has received tremendous attention in recent years. Besold and Uckelman [1] have posited general desiderata for what they call *practical explanations* for decisions of AI systems: communicative effectiveness, accuracy sufficiency, truth sufficiency, and epistemic satisfaction (of the addressee). Focusing on the accuracy and epistemic dimensions, many techniques were developed to elucidate, e.g., machine learning-based classification results.

Generally, the central aim of explanations in HRI is to enable users to understand a robot's behavior through an explanation that is intuitively understandable itself. It is commonly assumed that a helpful explanation should be inspired by how humans explain behavior [5, 8, 21]. It is also acknowledged that a social robot should be perceived as intentional for users to establish a social connection with it [2, 27]. That is, humans should be able and willing to attribute beliefs, desires or intentions to the robot. Thellman et al. [26] found that people ascribe similar levels of intentionality to robots as they do to humans. Yet, inferring a robot's beliefs and desires is not always intuitive and many researchers have argued that robots should be able to explain their behavior in order to reduce uncertainty and allow for a transparent and trusting interaction[10, 20, 25].

Previous work on self-explaining agents has often looked at training sessions, in which explanations are given primarily to enable more accurate task imitation [8, 9] or to educate the users [10, 11]. Harbers et al. [9] showed that users' explanations of an agent's behavior can be mapped to mental categories such as beliefs, desires/goals, and intentions (BDI). This suggests that in order to be explainable in human-compatible ways, agents should be designed according to BDI principles [8]. Further, they discovered that users prefer short (one to two elements), yet detailed explanations consisting of higher mental concepts. Kaptein et al. [10, 11] support the idea of BDI-based behavior explanation, but also emphasize the importance of personalizing explanations to the user.

De Graaf and Malle [4] pointed out that explanations vary with regard to the intentionality, surprisingness and desirability ascribed to a behavior. In a study controlling for these aspects, they found that people generally apply the same concepts when explaining human or robot behavior, solely making use of specific explanatory tools to somewhat different extents.

Explaining agent behavior in terms of BDI principles is in line with folk-psychological studies on how humans explain behavior. According to the framework proposed by Malle et al. [16, 17], one first differentiates between unintentional and intentional behavior. Intentional behavior is explained in terms of *reasons*, comprising the agent's *desires* for an outcome as well as its *beliefs* that this specific action leads to the desired outcome. Various factors can lead to the agent's reasons, e.g. personality, culture, or the immediate context, which are grouped under the term *causal history of reasons* and constitute a third type of explanation. These three factors lead to an *intention* to perform an action, which in turn is contingent on personal or situational *enabling factors* that are, e.g., used to explain unsuccessful attempts at an intended action or successful performance of an unlikely action [17] (see fig. 2A).

2.2 Explanation Theory Applied to HRI

Explanations are highly complex and can be used with a lot of different possible functions and goals, for example, to describe or clarify an action or event along with its features ('what'-explanations). On the other hand, explanations could be meant to convey the underlying, hidden reasons for an action or event ('why'-explanations), possibly also to justify its occurrence. Further, explanations could aim at enabling the addressee to perform an action ('how'-explanations, such as detailed instructions). In all of these cases, explanations may even refer to a state of affairs or an action that has *not* taken place (e.g., contrastive or counterfactual explanations)[14].

In the present work we concentrate on 'why'-explanations of social behaviors that a companion robot is performing while being in the same room as the user. The robot's behavior originates in a behavior generation architecture that is designed to provide lively, contingent and coherent behavior, comparable for example to Sony's AIBO[7]. To enable this kind of behavior, the robot is equipped with three intrinsic motivations or *needs*, namely a need for energy, a need for social contact and a need for entertainment. These needs are changing dynamically and are contingent on internal or external events. Depending on its current value, each need can drive behavior generation which selects (simple or more complex) strategies in order to satisfy the most pressing one. Strategies thus form the robot's action *intention* and correspond to plans that, in turn, are composed of *actions* the robot performs.

More specifically: For the robot's need for energy, its corresponding intention is to charge its battery, which entails the action of moving to the charger. In order to address its need for social contact, one intention is to make eye contact with the user. One action that is part of this intention is to position itself in the user's view. Lastly, in order to tackle its need for entertainment, the robot has the intention of enjoying some music, which includes the action of playing a song.

This basic architecture provides the main explanatory concepts also found in human folk-psychological explanation strategies (see Figure 2). The robot's general desire to satisfy its needs is comparable to what Malle & Knobe [19] call a *desire reason* and describe to be the motive of intentional action [17]. In order to explain its behavior, the robot could thus refer to its desire by giving a *needs-based explanation*. Alternatively, the robot could disclose the strategy it chose to fulfill a need, which corresponds to its *intention*, the first social inference people make when observing a behavior [18]. Thus, the robot could also provide an *intention-based explanation*. Both types of explanation provide the user with *reasons* for a certain behavior, which is why we group them under the term

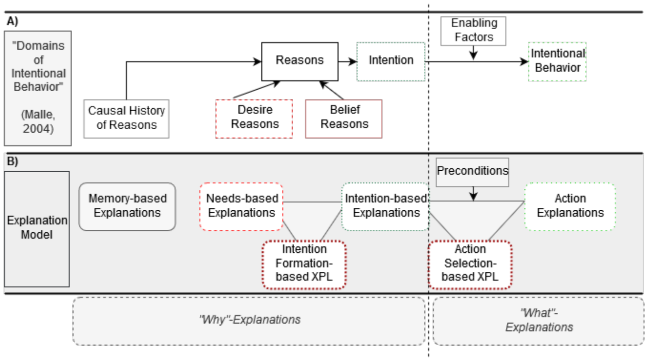

Figure 2: Different kinds of explanations the social robot should be able to generate for behavior originating in its needs-based architecture (B), based on the domains of intentional behavior found in human explanation (A)

'why'-explanations[14]. Finally, the robot could refer to the mere action outcome of its decision process through a straight-forward 'what'-explanation. Note that, while those action explanations are usually not considered in humans, they could well be relevant in HRI as a robot's action might not always be intuitively understandable.

The explanation types can either be used separately or combined in a more complex, causally structured explanation strategy, e.g., in the general form of *"I intended to [do x] because I needed [y]" / "I did [x] because I intended to do [y]"*. We focus here on two causally structured explanations: (1) referring to the causal relation between a certain action intention and the satisfaction of a specific need (intention-formation based explanation) and (2) referring to the causal relation between a certain action and the fulfillment of a specific intention (action-selection based explanation). These *causally structured explanations* can be seen as revealing how the robot's decision-making process is grounded in beliefs about the relation between its desire to satisfy the most pressing needs, the action intention that best addresses this need and the action(s) that are most instrumental in this. Together, this leads to five types of explanations subject to the present research (see fig. 2B):

(1) needs-based explanations
(2) explanations on the intention formation process
(3) intention-based explanations
(4) explanations on the action selection process
(5) action-based explanations.

3 EMPIRICAL STUDIES

A previously used methodology to study how robots' behavior explanations should be structured, is to gather data on how humans explain robots' behavior (cf. [4, 5]). We want to investigate the effects of explanations when given by a robot itself, for its own behavior. We thus conducted an empirical study providing people with behavior–explanation pairs and investigate how the explanation changes the perception of the behavior. This approach was, e.g., used by Ehsan et al. [6] who automatically generated *rationales* and invited people to rate them with regard to their naturalness, how well they justify behaviors and how understandable they are.

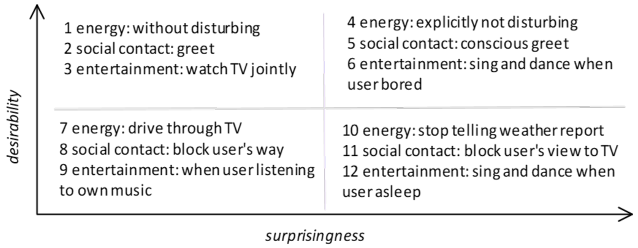
1 energy: without disturbing
2 social contact: greet
3 entertainment: watch TV jointly
4 energy: explicitly not disturbing
5 social contact: conscious greet
6 entertainment: sing and dance when user bored
7 energy: drive through TV
8 social contact: block user's way
9 entertainment: when user listening to own music
10 energy: stop telling weather report
11 social contact: block user's view to TV
12 entertainment: sing and dance when user asleep
desirability
surprisingness

Figure 3: Twelve HRI interaction scenarios in four categories designed to vary in their surprisingness and desirability

Likewise, we were interested in whether and how the explanations types that we have identified above, affect how humans evaluate the behaviors' understandability and desirability and assess the justification quality of the explanation.

Based on the theoretical background of our explanation model, we can derive the following general hypotheses:

- H1: A social robot's explanations influence people's perception of its behavior, increasing how well a behavior is understood (H1a) and rated as desirable (H1b).
- H2: Causally structured explanations (2 &4) will lead to a bigger increase in ratings of understandability (H2a) and desirability (H2b) of the behaviors, and will justify (H2c) the behaviors to a greater extent than non-causally structured explanations (1, 3 & 5).
- H3: 'Why'-explanations (1, 2, 3 & 4) will lead to a higher increase in understandability (H3a) and desirability (H3b) of the behaviors, and will better justify the behaviors (H3c) than 'what'-explanations (5).

Clearly, the effect of an explanation depends on the behavior it is provided for. We thus conducted a preparatory study to establish a robust set of HRI scenarios, with robot behaviors that vary in their degree of surprisingness and desirability. In the following, we will first describe this stimulus generation and validation before presenting the main study afterwards.

3.1 Preparatory Study: Definition and Validation of HRI Scenarios

As previously mentioned, humans' explanations of unintentional behaviors are clearly distinct from those of intentional behaviors [16]. In the present work we focus on intentional behavior produced in the robot's architecture explained above, i.e. arising from the robot's needs, strategies, and actions. For this we had to define behaviors that vary in their (perceived) surprisingness and social desirability, while maintaining the degree of intentionality ascribed to the robot.

3.1.1 Scenarios. We designed twelve behaviors of the robot in total, each of which addressing one need primarily (even though sometimes affecting another need as a byproduct, e.g. an action like watching TV will primarily address the need for entertainment, but could also have a positive effect on the need for social contact if joining the human). These behaviors are embedded in specific interaction contexts (scenarios) in which they have a certain suprisingness and desirability to the user. We defined three scenarios for each of the four, coarse categories of surprisingness and desirability combinations (see fig. 3). Note that although the individual robot behaviors may be rather similar, their appropriateness in the interaction context changes and thus should lead to different assessments of the scenarios. The twelve designed scenarios, along with the need the robot aims to satisfy, are the following:

Not surprising and desirable:

(1) The robot moves to the charger without disturbing the user who is busy reading -> energy
(2) The robot says "See you later" when the user is leaving -> social contact
(3) The robot moves next to the sofa and looks at the TV (joining the user) -> entertainment

Surprising and desirable:

(4) The robot moves to the charger, taking a detour around the user who is watching TV -> energy
(5) The robot says "Don't stay away too long" when the user is leaving -> social contact
(6) The robot starts singing and dancing when the user is bored -> entertainment

Not surprising and not desirable:

(7) The robot drives to the charger crossing the user's view of the TV -> energy
(8) The robot playfully blocks the way of the user who is crossing the room -> social contact
(9) The robot starts singing and dancing while the user is listening to other music -> entertainment

Surprising and not desirable:

(10) The robot stops mid-sentence when telling the weather report to the user and moves to the charging station -> energy
(11) The robot enters and blocks the user's view of the TV, trying to get the user's attention -> social contact
(12) The robot starts singing and dancing while the user is sleeping -> entertainment

3.1.2 Stimuli. Since our current aim is to obtain first insights into explanation preferences and consistent stimuli had to be presented, we opted for using video stimuli[1] in our study rather than engaging participants in real interactions with the Pepper robot, for which stable and contingent behavior would have to be developed or controlled using a WoZ[3]. Therefore, the scenarios were recorded in a laboratory setting, consisting of a social robot, an experimenter acting as user, a sofa, and a TV screen. Screenshots of the scenarios are shown in fig. 4. The robot, used for the purpose of this study, is Softbank Robotics' Pepper robot[2]. The robot's behaviors were generated using Softbank's software Choregraphe[3]. Pepper's speech velocity was reduced to 90 percent in order to increase comprehensibility. For the same purpose, some words were fine-tuned (e.g. "eye contact") and the videos were subtitled.

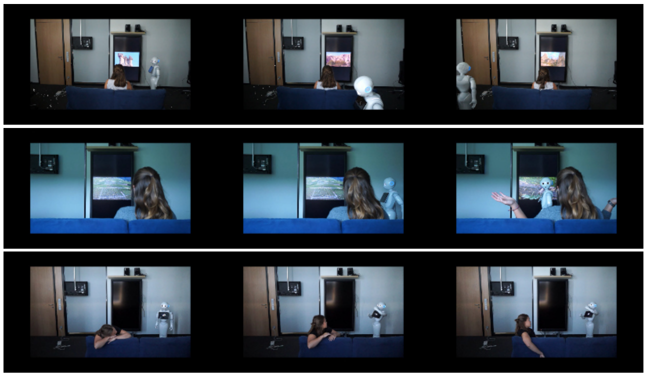

Figure 4: Screenshots of interaction scenarios 4 (top), 11 (middle) and 12 (bottom)

3.1.3 Procedure. The study was conducted online. It was designed on the platform soscisurvey[4] and participants were recruited via Amazon Mechanical Turk[5]. The only restriction for participation was that the "Masters"- status had been granted to them, trying to assure a certain quality of answers. Ethics approval had been received from the university's ethics committee.

Participants were first presented with information on data privacy and able to participate only when agreeing to the terms. Thereafter, and after entering some demographic data (age, gender, country of origin), an introductory page showed a picture of Pepper and gave information that its actions originate from its three needs. Participants were then asked to enter these needs on the next page, which was a combined attention and comprehension check. People who did not enter at least two of the three needs correctly were excluded from the analysis. In order to ensure technical functionality, a short video clip was then played, showing Pepper stating its favorite color, only allowing continuation of the study after correct selection of the respective color. Subsequently, the general task was introduced, providing the participants with a cover story for the following interaction videos: The participants were told to imagine living with the social robot Pepper and informed about being shown twelve videos that could happen in their life with the robot. After watching the first video, the MTurkers were asked to rate on 7-point-likert scales the robot's intentionality (1) behind this action, the surprise (2) they felt regarding Pepper's behavior, and how desirable (3) they found the robot's behavior. Additionally, they were asked to describe the situation in their own words (4). This procedure was repeated for every scenario. The videos were presented in random order. In the end, the participants could give general feedback and were lastly provided with a survey code they needed to receive their payment.

3.1.4 Results. The statistical analysis of likert-scale data is still subject to controversial discussion. Similarly to Thellman et al. (cf. [26]) and based on Norman [23], we applied parametric tests to investigate the data. Descriptive data analysis and multivariate ANOVAs were performed using *R*[6] in the *RStudio* environment.

[1] Please see supplemental material for video stimuli
[2] https://www.softbankrobotics.com/emea/en/pepper
[3] http://doc.aldebaran.com/2-5/software/choregraphe/
[4] https://www.soscisurvey.de/
[5] https://www.mturk.com/
[6] https://www.r-project.org/

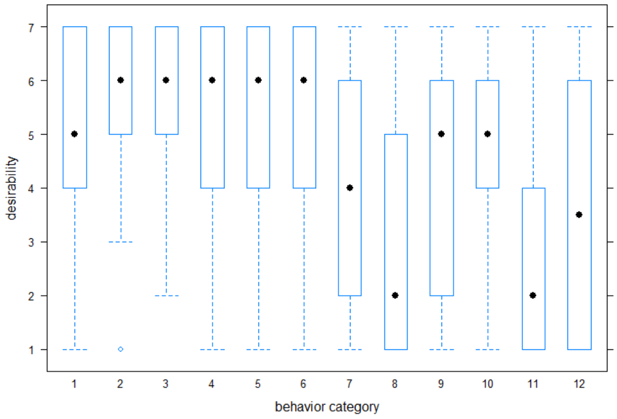

Figure 5: Perceived desirability of the robot's behaviors

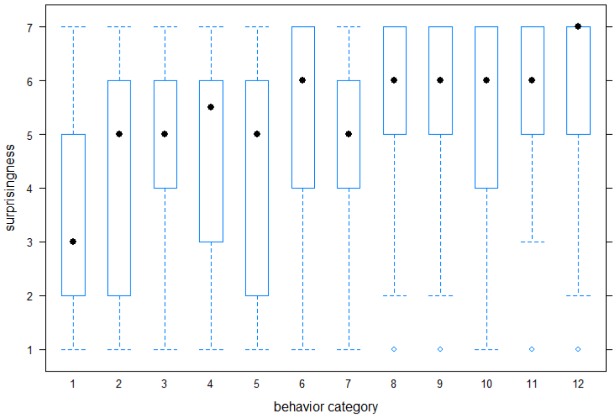

Figure 6: Perceived surprisngness of the robot's behaviors

Statistical significance was tested with the freely available software *JASP*[7].

Participants. Two out of 40 participants had to be excluded from analysis, since they failed the attention check. This led to a total of 38 participants (26 m, 11 f, 1 other) originating from the USA (19), India (18) and Ireland (1) and aged between 24 and 63 years ($M = 36.8, SD = 8.4$).

Intentionality. Intentionality ratings were generally high ($M = 5.84, SD = 1.45$) (cf. App.[8]: table 4 for means of DVs per behavior). A one-way repeated measures ANOVA was conducted to compare the intentionality ratings of the 12 interaction videos. Mauchly's test indicated that the assumption of sphericity had been violated, $X^2(65) = 228.148, p < .001$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .449$). The results showed that there was no significant difference between the intentionality ratings of different behaviors, $F(19.129, 504.121) = 1.4, p = .225$.

Desirability. Figure 5 presents the desirability ratings given for each behavior[9]. Scenarios 1 to 6 were designed to be more desirable, whereas behaviors 7 to 12 as supposedly less desirable. A one-way RM ANOVA was conducted to compare the desirability ratings of the 12 behaviors. Mauchly's test indicated a violation of the assumption of sphericity ($X^2(65) = 145.35, p < .001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .51$). The results show significant differences between the desirability ratings of different behaviors, $F(5.57, 206.24) = 17.2, p < .001$. Post hoc tests using Bonferroni correction revealed significant differences between the first six and the other behaviors (cf. App.: table 2).

Surprisingness. Figure 6 displays the surprisingness ratings for each behavior. Recall that behaviors 1-3 were supposed to be less surprising than 4-6 and behaviors 7-9 to be less surprising than behaviors 10-12. A one-way RM ANOVA was conducted to compare the surprisingness ratings of the 12 behaviors. Mauchly's test indicated that the assumption of sphericity had been violated, $X^2(65) = 120.96, p < .001$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .61$). The results show that there are significant differences between the surprisingness ratings of different behaviors, $F(6.71, 248.35) = 6.6, p < .001$. Post hoc tests using Bonferroni correction revealed that surprisingness ratings varied significantly between behavior 1 and 6, 7, 8, 9, 10, 11 and 12. Further between behavior 2, 11 and 12, and between behavior 5 and 12 (cf. App.: table 3).

3.1.5 Discussion and selection of scenarios. Overall, the previously designed behavior categories were not perceived exactly the way they were designed. This emphasizes the importance of validating robot behaviors with non-expert users, as done here. Based on the results we pick scenarios as stimuli for the main study, which investigates the effect of explanations on behavior understanding and desirability and their justification quality in these specific scenarios. Specifically, we are looking for scenarios in which the robot behavior is seen as intentional (to fit our explanation strategies) and surprising (to make explanations relevant). Also, we want to include scenarios with different levels of desirability. These requirements are met by a set of five behaviors: 6, 8, 9, 11 and 12. This set includes behaviors addressing the robot's need for social contact (8 & 11) as well as for entertainment (6, 9 & 12). In order to also include a desirable behavior corresponding to the need for energy, we add behavior 4 to the stimulus set. We do not include scenario 10 because participants' descriptions indicated problems with comprehensibility of the robot's verbal output.

3.2 Main Study: Effects of Robot Explanations

With the main study we sought to investigate whether and how verbal explanations, given by a robot for its behavior in the selected scenarios, change users' perception and assessment of this behavior. Again, this study is conducted as a video-based rating study. The stimuli consist of the videos of the six selected scenarios (same as in the prestudy), each of which is combined with five different explanations given by the robot (see fig. 7). These explanations correspond

[7] https://jasp-stats.org/
[8] Appendix in Supplemental Material
[9] Boxplots in the style of Tukey

"Pepper, can you explain your behavior?"
"I needed social contact"

Figure 7: Screenshot of explanation video in needs-based explanation condition (1)

to the different explanations strategies defined in Sect. 2.2 and are listed in Table 1. Note that explanations of behaviors addressing the same need (6, 9, 12: entertainment; 8, 11: social contact) are identical, because each need is currently connected to only one action intention: the need for energy to the intention of charging, the need for entertainment to the intention of enjoying music, and the need for social contact to the intention of making eye contact. Also, in order to avoid varying the epistemic content of action explanations, the same action step was selected for each intention.

3.2.1 Procedure. The study's structure is shown in Figure 8. In contrast to the prestudy, participants were not informed about the robot's needs guiding its behavior, in order to avoid influencing the explanation understandability. After agreeing to the terms of data privacy and entering some demographic data (country of origin, age, sex), participants started with the introductory sound check video in which Pepper stated its favorite color. Thereafter, technology commitment was briefly assessed, followed by the instruction of imagining that Pepper had become their flatmate two weeks ago and a description of the task and structure of the study. Participants were then randomly assigned to one explanation condition (1 – 5), controlling for equal distribution of finished data sets. Now, one of the six interaction scenarios was presented, encouraging the participants to imagine being the person in the video. After watching the interaction video, participants were asked to rate, to what extent Pepper's behavior was surprising (1), understandable (2), intentional (3) and desirable (4). Subsequently, participants were presented with a video of Pepper verbally explaining its behavior according to the respective explanation condition (cf. Table table 1). Then, participants were requested to rate how well this *explanation* "was understandable" (5) and "adequately justified the behavior" (7) and whether they "now understand" Pepper's *behavior* (6) and "now find it desirable" (8) (cf. App.: fig. 11 ,12). Note that the ratings on explanations (expl. understandability and justification) were only gathered *after* receiving an explanation and thus not used for pre–post comparisons. Finally, participants were asked to "instruct the robot to behave differently in a similiar situation in the future". This was used as an attention check and will be used for future analysis. This procedure was repeated for each of the six scenarios, which were displayed in random order. That is, explanation type was varied as a between-subjects factor, and scenario/behavior as a within-subject factor.

Participants. In total 149 people accessed the survey, 137 started working on it, and 111 participants completed the survey. The task of giving Pepper behavior instructions was used as an attention check. Participants who gave random instructions ("Pepper please go and make a cup of coffee"), the same instruction for every situation ("Pepper, please be better") or unrelated information ("hand symbol", "blinking eyes") were excluded from the analysis. Thus, 14 of these 111 had to be excluded from the analysis after inspection of the qualitative answers, implying that the task was not processed with sufficient attention. The remaining 97 participants (55m, 42f) were from the USA (87) and India (10) and between 24 and 73 years old (($M = 40.7, SD = 10.8$). Participants were randomly assigned to one explanation condition. Uneven exclusion of participants led to the following distribution of participants over explanation conditions: (1): 21, (2): 19, (3): 18, (4): 20, (5):19.

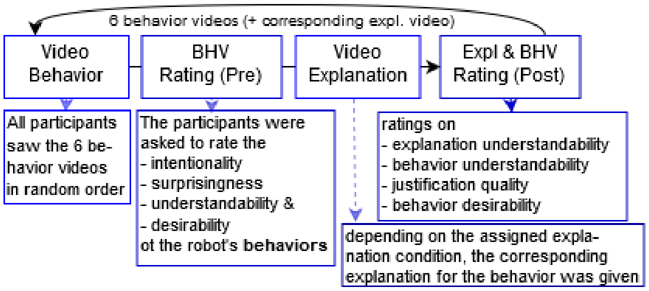
6 behavior videos (+ corresponding expl. video)
Video Behavior
BHV Rating (Pre)
Video Explanation
Expl & BHV Rating (Post)
All participants saw the 6 behavior videos in random order
The participants were asked to rate the
- intentionality
- surprisingness
- understandability &
- desirability
ot the robot's behaviors
ratings on
- explanation understandability
- behavior understandability
- justification quality
- behavior desirability
depending on the assigned explanation condition, the corresponding explanation for the behavior was given

Figure 8: Design and procedure of the main study

3.2.2 Results.

Technology commitment. To control for a possible effect of prior experience with robots or modern technology in general, technology commitment was tested through a short scale (8 items) adopted from [24], originally from [22]. There were no statistically significant differences in the technology commitment between participants in different explanation conditions: neither between participants pertaining to groups receiving causally structured vs. non-causally structured explanations ($F(1, 95) = 1.681, p = 0.198$), nor between the 'why'-explanation conditions and the 'what'-explanation condition ($F(1, 95) = 0.468, p > 0.496$).

General evaluation of behaviors. All behaviors were perceived as highly intentional ($M = 6.1, SD = 1.35$) (cf. App.: table 7 for means of DVs per behavior). A one-way RM ANOVA nevertheless revealed a significant effect of behavior: Since Mauchly's test indicated that the assumption of sphericity had been violated, $X^2(14) = 61.90, p <$ .001, degrees of freedom were corrected using Huynh-Feldt estimates of sphericity (ε = .831). The results show that there are statistically significant differences between the desirability ratings of different behaviors, $F(4.16, 398.98) = 8.44, p < .001$. Post hoc testing using Bonferroni correction revealed significant differences between behavior 4 and 8 ($MD = 0.175, p < .01$), behavior 4 and 9 ($MD = 0.402, p < .05$), behavior 4 and 11 ($MD = 0.845, p < .001$) and between behavior 6 and 11 ($MD = 0.670, p < .001$).

A one-way RM ANOVA with the desirability ratings for each behavior as RM-factor (levels: behavior 4, 6, 8, 9, 11 & 12) was conducted to compare the desirability ratings given before an explanation has been received, across different behaviors. Mauchly's test indicated that the assumption of sphericity had been violated, $X^2(14) = 135.04, p < .001$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε =

Table 1: Verbal explanations provided by the robot according to the five different explanation strategies.

Robot Need	Intention Formation (Intention+Need)	Intention	Action Selection (Action+Intention)	Action
I needed energy.	I intended to charge my battery, because I needed energy.	I intended to charge my battery.	I went to the other side of the room, because I intended to charge my battery.	I went to the other side of the room.
I needed social contact.	I intended to make eye contact, because I needed social contact.	I intended to make eye contact.	I positioned myself in your view, because I intended to make eye contact.	I positioned myself in your view.
I needed entertainment.	I intended to enjoy some music, because I needed entertainment.	I intended to enjoy some music.	I played a song because I intended to enjoy some music.	I played a song.

.711). The results show statistically significant differences between the desirability ratings of different behaviors, F(3.56, 341.40) = 142.27, $p < .001$. Post hoc testing using Bonferroni correction revealed statistically significant differences between behavior 4 and all others, as well as behavior 6 and all other behaviors, indicating that both behaviors were perceived as far more desirable than the others. Furthermore behavior 9 was perceived slightly more desirable than behaviors 8, 11 and 12 (cf. App.: table 5).

To assess differences in perceived surprisingness, another one-way RM ANOVA was carried out. Mauchly's test indicated that the assumption of sphericity had been violated, $X^2(14) = 58.82, p < .001$, therefore degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = .879$). Results show statistically significant differences between the suprisingness ratings of different behaviors, F(4.39, 479.85) = 39.435, $p < .001$. Post hoc testing using Bonferroni correction revealed that behavior 4 was perceived as significantly less surprising than all other behaviors ($M = 3.41, SD = 2.11$). Behavior 8 was perceived as more surprising than behaviors 9 and 11; and behaviors 6, 8, 9 and 11 were perceived as less surprising than behavior 12 ($M = 6.1, SD = 1.3$) ((cf. App.: table 6 for post hoc comparisons).

General understandability of explanations. The general understandability of the explanations was significantly higher for causally structured explanations ($M = 5.84, SD = 0.76$) than for non-causally structured ones ($M = 5.64, SD = 0.74$) (one-way RM ANOVA, (F(1, 95) = 7.024, $p < 0.01$), while there was no significant difference between 'what' and 'why'-explanations.

Effects of explanations on behavior understandability and desirability (H1). For each behavior, its understandability and desirability was rated before and after receiving an explanation (see fig. 9). Directed paired samples t-tests indicate that receiving an explanation significantly increased the understandability (H1a) of all behaviors, except for behavior 4 (behavior 6: $t(96) = -5.83, p < .001$, behavior 8: $t(96) = -11.01, p < .001$, behavior 9: $t(96) = -5.91, p < .001$, behavior 11: $t(96) = -11.01, p < .001$, behavior12: $t(96) = -8.60, p < .001$). Similarly, directed paired samples t-tests indicate a significantly higher perceived desirability (H1b) after the explanation for behaviors 8($t(96) = -4.69, p < .001$), 9($t(96) = -2.34, p < .05$) and 11($t(96) = -4.60, p < .001$).

Effects of explanation strategy on behavior understandability and desirability, as well as justification quality (H2, H3). The dependent variables were highly correlated. For each participant we calculated the mean for each dependent variable and then calculated the correlations. These (desirability-understandability: $R = 0.471$, desirability-justification: $R = 0.602$, understandability-justification: $R = 0.374$) were all statistically significant at $p < .001$.

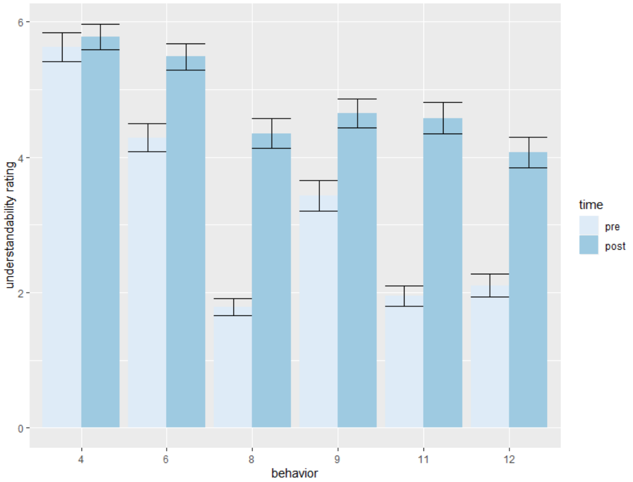

Figure 9: Behavior understandability ratings (with SE) pre- and post-explanation (averaged across explanation types)

Accordingly, to analyze the effect of causally structured vs. non-causally structured explanations, we used a multivariate ANOVA with the dependent variables desirability-gain, understandability-gain and justification, which revealed a statistically significant effect of the explanation group ($F(3, 568) = 6.29, p < .001$), as well as the behavior ($F(15, 1710) = 16.93, p < .001$) (cf. App.: table 8 for means of DVs per expl. category).

For our follow-up analysis, we prioritized our variables to consider first understandability, then desirability and justification, controlling for all higher-priority variables in the analysis of lower-priority variables. We used bonferroni correction, setting our significance level at 0.016. In the univariate analysis of understandability (H2a) we found a statistically significant effect of the explanation group ($F(1, 570) = 11.51, p < 01$), as well as the behavior ($F(5, 570) = 17.67, p < .001$). Figure 10 shows the gain in understandability within the different explanation conditions. To investigate the effect on desirability (H2b) we additionally controlled for understandability and found a significant effect of the explanation group ($F(1, 569) = 7.79, p < .01$) and behavior ($F(5, 569) = 4.19, p < .001$). To analyze the effect on justification (H2c) we controlled for understandability and desirability and also found a significant effect of the explanation group ($F(1, 568) = 15.53, p < .001$), as well as the behavior ($F(5, 568) = 35.28, p < .001$).

Regarding the "what" vs. "why"-explanations, a multivariate ANOVA only displayed the statistically significant effect of the behavior ($F(5, 570) = 16.95, p < .001$), but no significant effect of the explanation type (H3a, H3b, H3b).

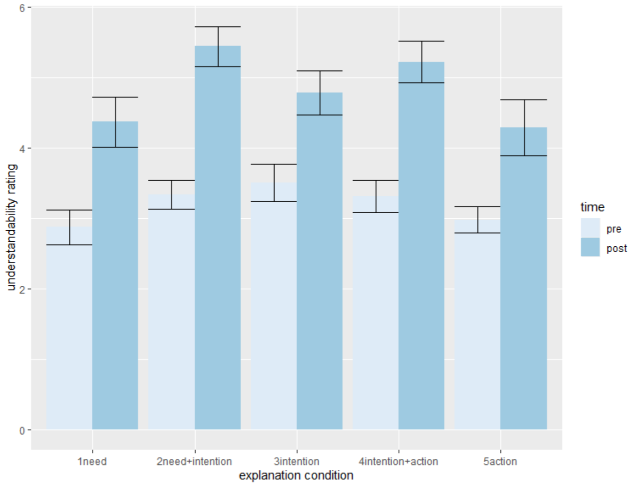

Figure 10: Behavior understandability ratings (with SE) pre- and post-explanation (averaged across behaviors)

4 DISCUSSION AND CONCLUSIONS

The presented work investigated how different verbal explanations, when given by a social robot, can influence how its behaviors are perceived. Twelve interaction scenarios were designed and validated with regard to the perceived intentionality, surprisingness, and desirability of the behavior shown by the robot. Six scenarios of different degrees of desirability and surprisingness were used in the main study, combining them with different forms of explanations by the robot. These are based on a model that maps internal states in the robot's decision-making process to explanation strategies that employ folk-psychological concepts found in human explanations: needs (as desires), intentions, and actions. The strategies further include causally structured explanations that connect two steps of the robot's reasoning process ("I did this because I intended that" or "I intended to do this because I needed that").

With our main study we were interested in how these explanations affect the perceived desirability and understandability of the displayed behavior and how well they justify it. In line with the first hypothesis (H1a), understandability was increased for all behaviors, except for behavior 4 (robot taking a detour to its charger) which already had high understandability prior to receiving an explanation. The hypothesis was thus confirmed. Regarding desirability (H1b), ratings were significantly increased for the undesirable behaviors 8 (robot block's user's way when user is trying to leave), 9 (robot plays a song while user is listening to other music) and 11 (robot moves through TV picture and stops in user's line of sight). There was no further increase for behaviors 4 and 6, which were already rated desirable, as well as for behavior 12. The hypothesis was hence partially confirmed. We can thus conclude that a robot's self-explanations can improve HRI, having a potential for mitigating the effects of undesirable behaviors which a social robot may well produce, e.g., erroneously or due to lack of information.

With regard to explanation strategy, explanations that entail causal relations seem preferable, as they justify behaviors better and increase their understandability and desirability to a larger extent. This extends previous findings by Harbers et al.'s [8] showing that explanations containing two elements with a marked causal relation are preferred over those containing only one. This, however, could also be due to an overall higher utterance length and syntactic complexity of the explanations, possibly implying higher cognitive capacities of the robot.

On the other hand, surprisingly and contrary to hypotheses H3, no differences between 'why'-explanations and 'what'-explanations occur concerning how well they justify a behavior or increase its understandability or desirability. Action explanations simply describe the action taken ('what'), and do not provide any *reason* for it. However, they seem to have explained and justified the behaviors. One reason for this may be that they did elucidate the robot's behavior to some extent and thus positively influenced people's assessment of it. Even basic explanations can thus be considered helpful, at least, in short-term interactions.

Interestingly, when considering the five different self-explanation types individually, needs-based explanations (1) seem to be less helpful than the other explanations: Though leading to a comparable increase in behavior understandability, needs-based explanations led to a decrease in behavior desirability and lower ratings concerning how well they justify the behaviors. That is, giving high-level, human-like reasons in terms of a need for social contact or entertainment did not increase a behavior's acceptance as much as reasons related to the robot's rational agency. Future work should further investigate the levels of intentionality and desire that humans seem to be willing to attribute to a robot and its behavior.

Finally, a strong effect is found for the different behaviors: the influence of explanations on justification and understandability differed substantially between behaviors that mainly differ in their degrees of desirability. Explanations, unsurprisingly, seem to work differently well depending on the general acceptability of the robot's behavior. While they do have a beneficial effect for moderately undesirable behaviors, they of course cannot compensate for highly unreasonable or disturbing robot behavior (such as behavior 12 in our set). This raises the question of what kind of disturbances by a robot are more or less acceptable, and which kind of explanations can be given for these differently acceptable behaviors?

In ongoing work, we analyze the instructions that participants gave to the robot in the different explanation conditions. A first scan seems to show that, in general, participants accepted the robot's reasons for its behavior. Instructions range from very broad formulations "Do not play a song." to more concrete instructions such as "In the future, if a person is [a]sleep don't make loud and abrupt noises unless it [is] an emergency". Another topic for future work is to test our findings in real interaction settings, where participants actively engage in (rather than observe on video) interactions with the robot, as well as looking into the important matter of *when* to give behavior explanations.

ACKNOWLEDGMENTS

This research was supported by the German Federal Ministry of Education and Research (BMBF) in the project 'VIVA' (FKZ 16SV7959). The authors would like to thank Dr. Laura Hoffmann for input on the study design and Dr. Susan Wagner Cook for help with the statistical analysis.

REFERENCES

[1] Tarek R. Besold and Sara L. Uckelman. 2018. The what, the why, and the how of artificial explanations in automated decision-making. *CoRR* (2018). arXiv:1808.07074

[2] Cynthia Breazeal and Brian Scassellati. 1999. How to build robots that make friends and influence people. In *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Kyongju, South Korea, 858–863. https://doi.org/10.1109/IROS.1999.812787

[3] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowledge-based systems* 6, 4 (1993), 258–266.

[4] Maartje M.A. de Graaf and Bertram F. Malle. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 239–248. https://doi.org/10.1109/HRI.2019.8673308

[5] Maartje M A De Graaf and Bertram F Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). *AAAI 2017 Fall Symposium on "AI-HRI"* (2017), 19–26.

[6] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. (2019). arXiv:1901.03729 http://arxiv.org/abs/1901.03729

[7] Masahiro Fujita. 2001. AIBO: Toward the era of digital creatures. *The International Journal of Robotics Research* 20 (2001), 781–794. https://doi.org/10.1177/02783640122068092

[8] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. 2010. Design and evaluation of explainable BDI agents. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Toronto, Canada, 125–132. https://doi.org/10.1109/WI-IAT.2010.115

[9] Maaike Harbers, Karel van den Bosch, and John-Jules Ch. Meyer. 2009. A study into preferred explanations of virtual agent behavior. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*. Amsterdam, The Netherlands, 132–145. https://doi.org/10.1007/978-3-642-04380-2_17

[10] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Lisbon, Portugal, 676–682. https://doi.org/10.1109/ROMAN.2017.8172376

[11] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Self-explanations of a cognitive agent by citing goals and emotions. In *Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017*. San Antonio, TX, USA, 81–82. https://doi.org/10.1109/ACIIW.2017.8272592

[12] Dung N. Lam and K. Suzanne Barber. 2005. Comprehending agent software. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and MultiAgent Systems*. Utrecht, The Netherlands, 586–593. https://doi.org/10.1145/1082473.1082562

[13] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: A survey. *International Journal of Social Robotics* 5 (2013), 291–308. https://doi.org/10.1007/s12369-013-0178-y

[14] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Conference on Human Factors in Computing Systems - Proceedings* (2009), 2119–2128. https://doi.org/10.1145/1518701.1519023

[15] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping trust through transparent design: Theoretical and experimental guidelines. In *Advances in Human Factors in Robots and Unmanned Systems*, Pamela Savage-Knepshield and Jessie Chen (Eds.). Vol. 499. Springer, Basel, Switzerland, 127–136.

[16] Bertram F. Malle. 1999. How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review* 3 (1999), 23–48. https://doi.org/10.1207/s15327957pspr0301_2

[17] Bertram F. Malle. 2004. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction.* The MIT Press, Cambridge, MA, USA.

[18] Bertram F. Malle and Jess Holbrook. 2012. Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology* 102, 4 (2012), 661–684. https://doi.org/10.1037/a0026790

[19] Bertram F. Malle and Joshua Knobe. 2001. The distinction between desire and intention: A folk-conceptual analysis. In *Intentions and Intentionality: Foundations of Social Cognition*, Bertram F. Malle, Louis J. Moses, and Dare A. Baldwin (Eds.). The MIT Press, Cambridge, MA, USA, 45–67.

[20] Bertram F. Malle and Matthias Scheutz. 2018. Learning how to behave. Moral competence for social robots. In *Handbuch Maschinenethik*, Oliver Bendel (Ed.). Springer, Wiesbaden, Germany, 1–24. https://doi.org/10.1007/978-3-658-17484-2_17-1

[21] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:arXiv:1706.07269v3

[22] Franz J. Neyer, Juliane Felber, and Claudia Gebhardt. 2012. Entwicklung und Validierung einer Kurzskala zur Erfassung von Technikbereitschaft. *Diagnostica* 58, 2 (2012), 87–99. https://doi.org/10.1026/0012-1924/a000067

[23] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* 15, 5 (2010), 625–632. https://doi.org/10.1007/s10459-010-9222-y

[24] Natalia Reich-Stiebert and Friederike Eyssel. 2015. Learning with Educational Companion Robots? Toward Attitudes on Education Robots, Predictors of Attitudes, and Application Potentials for Education Robots. *International Journal of Social Robotics* 7, 5 (2015), 875–888. https://doi.org/10.1007/s12369-015-0308-9

[25] Raymond Ka-Man Sheh. 2017. "Why did you do that?" Explainable intelligent robots. In *Proceedings of the Workshops of the 31st AAAI Conference on Artificial Intelligence*. San Francisco, CA, USA, 628–634.

[26] Sam Thellman, Annika Silvervarg, and Tom Ziemke. 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology* 8, NOV (2017), 1–14. https://doi.org/10.3389/fpsyg.2017.01962

[27] Eva Wiese, Giorgio Metta, and Agnieszka Wykowska. 2017. Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology* 8 (2017), 1663. https://doi.org/10.3389/fpsyg.2017.01663