Warning Signals for Poor Performance Improve Human-Robot Interaction

Rik van den Brule
Donders Centre for Brain, Cognition, and Behaviour, Radboud University
Behavioural Science Institute, Radboud University
Gijsbert Bijlstra
Behavioural Science Institute, Radboud University
Ron Dotsch
Department of Psychology, Utrecht University
Pim Haselager
Donders Centre for Brain, Cognition, and Behaviour, Radboud University
Daniël H. J. Wigboldus
Behavioural Science Institute, Radboud University

The present research was aimed at investigating whether human-robot interaction (HRI) can be improved by a robot's nonverbal warning signals. Ideally, when a robot signals that it cannot guarantee good performance, people could take preventive actions to ensure the successful completion of the robot's task. In two experiments, participants learned either that a robot's gestures predicted subsequent poor performance, or they did not. Participants evaluated a robot that uses predictive gestures as more trustworthy, understandable, and reliable compared to a robot that uses gestures that are not predictive of their performance. Finally, participants who learned the relation between gestures and performance improved collaboration with the robot through prevention behavior immediately after a predictive gesture. This limits the negative consequences of the robot's mistakes, thus improving the interaction.

Keywords: human-robot interaction, error prevention, predictive signals, evaluation

Introduction

Robotic autonomous systems are becoming more common in everyday life. Successful robotic applications must be capable of operating in versatile, dynamic environments, and of interacting with people present. As such, they can be considered socially interactive robots (Fong, Nourbakhsh, & Dautenhahn, 2003). The uncontrolled domestic environment they operate in can change quickly, and a task that is relatively simple in one situation might become increasingly complex in other situations. For example, an otherwise simple task for a butler robot such as serving a glass of water may become quite difficult when its tactile sensors are dirty and it can no longer feel how well it holds the glass. Likewise, handing over a glass of beer may be easy in a situation where the receiver is alone and sitting quietly behind a table but much more complicated once a party has started and the receiver is found dancing on a crowded floor.

It is impossible for robot designers to foresee every scenario and to create a robot that can function reliably under all circumstances. A robot that can respond to the inherent uncertainty and changing nature of the world seems to be the next best thing. We suggest that by indicating a level of confidence or uncertainty about their actions, robots could calibrate a human's expectations regarding a robot's performance. This may increase the level of trust in a robot's capacities, as well as solicit useful interventions to improve task performance.

Some prior work in human-robot interaction (HRI) and ergonomics exists in which uncertainty of automated systems is manipulated. For instance, displaying uncertainty in an electronic vehicle's range estimation has been shown to improve driving experience and behavior (Jung, Sirkin, Gür, & Steinert, 2015). Other work demonstrated that a robotic advisor is perceived as more considerate when its advice contains hedges and discourse markers, parts of speech that contain uncertainty statements (Torrey, Fussell, & Kiesler, 2013). Such a strategy might also apply for socially interactive robots. Consistent with this, Moon, Panton, Van der Loos, and Croft (2010) suggested that using human hesitation gestures may let a robot communicate uncertainty through nonverbal communication channels. As yet, we believe that no research empirically tested this proposition and its consequence for interaction in an HRI context.

In many situations where the outcome of a robot's task is uncertain, minor interventions by humans may ensure the successful completion of the robot's task. In the butler robot example above, someone might reach for the glass more quickly or with more attention when he or she notices that the robot is not holding the glass as firmly as it usually does. People who have experience with a robot may perform these helpful acts because they are familiar with the limits of the robot system (Sung, Grinter, & Christensen, 2010). However, it takes time to develop this knowledge. Therefore, someone who is unfamiliar with the capabilities of a robot may not know when a failure is likely to happen, and the HRI may suffer as a consequence. In principle, any sort of signal can be used to indicate that a robot is uncertain in completing a task. For instance, Tellex et al. (2014) showed that a robot that solicits help with a specific request when a failure condition is encountered increases task performance. Other work has shown that using movements and vocal cues can be used effectively to convey the purpose and required assistance of a minimalistic trash can robot (Yamaji, Miyake, Yoshiike, de Silva, & Okada, 2011). We propose that a humanoid robot might capitalize on existing mechanisms of human signaling, such as body language, to facilitate the understanding of such signals (e.g., by making a gesture that is used naturally by humans to display uncertainty, such as scratching one's forehead).

Nonverbal communication is an important form of natural interaction between humans (Feldman & Rimé, 1991) and has been shown to influence social interactions (Ekman, Friesen, & O'Sullivan, 1988). Recent research indicates that similar results are possible in HRI (Stanton & Stevens, 2014). We consider nonverbal behavior to be an important aspect of robotics specifically because of a robot's embodiment. That is, a robot's physical presence enables new pathways to express its internal state. Especially when a humanoid robot is concerned, which can be anthropomorphized, body language (such as hand gestures) could yield novel insights in the fields of HRI and human factors.

In the current contribution, we suggest that gestures made by a humanoid robot could function as a signal of uncertainty by alerting users to impending failures. These signals can have a positive effect on the evaluation of, and collaboration with, the robot, because it enables people to prevent or mitigate the robot's mistake. One of the main factors in interpersonal evaluation is trust (DeSteno et al., 2012; Schlenker, Helm, & Tedeschi, 1973). There is evidence that trust can be interpreted as a general positive-negative judgment in interpersonal evaluation (Oosterhof & Todorov, 2008). Recent work has shown that trust also plays a large role in HRI (for an overview, see Hancock et al., 2011). Therefore, we expect trust to be affected by the predictability of a robot's performance by its gestures, because the interaction and the robot itself will be perceived as more pleasant. We will present two experiments that aim to provide empirical support for the working of such a mechanism.

Contingency Learning

The capacity to learn without explicit instruction that certain signals (such as gestures) provide information about the likelihood of certain events (such as mistakes) is a psychological process called contingency learning (for an overview, see Shanks, 2007). Early studies on classical conditioning indicated that people and animals alike can learn the pairing of an unconditioned stimulus (US), such as receiving food, with a conditioned stimulus (CS), such as a ringing bell (Pavlov, 1927). The association between the CS with the US can become so strong that natural responses (such as salivating) are expressed even when the CS is presented without the US. Similarly, once people associate a robot gesture (CS) with a bad performance, such as making a mistake (US), their response (intervention to limit the consequences of the mistake) may be expressed even when the CS is presented without the US.

Contingency learning works regardless deviations from a strict contiguity of the CS with the US (i.e., whether the CS occurs at the same time as the US), as shown by Rescorla (1968). When a CS is presented before the US (forward conditioning), people will prepare for the US. Also, it is not necessary for a CS to be

perfectly contingent on a US, as long as its appearance coincides with the US above chance level. As a result, the robot does not have to be flawless in predicting its task performance for these signals to be effective, which makes the use of contingency learning especially useful as a mechanism for improving HRI.

Human-Robot Trust

Gestures indicating upcoming poor performance may increase a person's trust in a robot, besides improving the outcome of the HRI. From a psychological perspective, trust is seen as an important variable that mediates a human's acceptance of technology (Muir, 1994; Parasuraman & Riley, 1997). Trust is described as the willingness to be vulnerable to someone else's actions in a situation characterized by risk and uncertainty (Mayer, Davis, & Schoorman, 1995). As such, trust is often used in research on trust in automation (Lee & See, 2004) and, by extension, HRI (Sanders, Oleson, Billings, Chen, & Hancock, 2011).

Previous research has shown that trust in robots is mainly affected by a robot's performance, although other robot attributes also play a role (Hancock et al., 2011). For example, Van den Brule, Dotsch, Bijlstra, Wigboldus, and Haselager (2014) provided initial evidence that a robot is trusted more when its behavioral style (e.g., shaky movements) matches its performance (making a mistake). Here we will follow up on our earlier study by examining our earlier suggestion (Van den Brule, Bijlstra, Dotsch, Wigboldus, & Haselager, 2013) that a robot that nonverbally signals uncertainty in a way that is contingent upon its actual performance will help to calibrate a user's trust in the robot.

The Present Work

In the present research, we specifically focused on nonverbal and simple signals, which can be learned implicitly by human users (that is, without explicit instruction and without necessarily being fully aware of what is learned) during a relatively short interaction. Based on contingency learning principles, we hypothesized that users can learn without instruction whether a robot's gestures are predictive of poor performance. Moreover, we hypothesized that when a user notices the predictive value of such a signal, he or she will take pre-emptive actions to ensure the robot will continue to perform well, or to prevent the potential consequences in case the robot makes a mistake. This preventive behavior should occur only when the robot's signal, such as a gesture, is relatively contingent on the robot's mistakes. Moreover, we predicted that a robot equipped with such behavior is trusted more, because users can understand the system better and evaluate it as more reliable (Madsen & Gregor, 2000). Understandability, predictability, and reliability (components of trust; e.g., Jian, Bisantz, & Drury, 2000; Lee & See, 2004; Sheridan, 1988) can be measured with questionnaires concerning the user's perception of various aspects of the robot's behavior (for details see *Study 1, procedure*, below). Additionally, by means of path analyses, we explored whether the learned contingency affected participants' behavior during the interaction with the robot and the extent to which that influences their evaluation of the robot.

Studies 1 and 2 were designed to investigate whether participants were able to learn the contingency between a robot's behavioral gestures and its performance level. Furthermore, we measured whether learning this contingency affected participants' trust-related evaluations. In Study 1, a video-HRI paradigm (VHRI; Dautenhahn, 2007; Syrdal, Koay, Gácsi, Walters, & Dautenhahn, 2010) with a simulated Nao robot was used to measure the contingency learning of the participants by asking them to predict the robot's actions directly after each observation of the robot. Afterward, we measured participants' evaluation of the robot's trustworthiness, reliability, and understandability. In Study 2, participants interacted with a real Nao robot, which allowed us to observe the effects of the learned contingency on participants' spontaneous and unconstrained intervening behavior. In doing so, Study 2 served to extend findings from the less realistic but more controlled video paradigm of Study 1 (and studies presented in Van den Brule et al., 2013) to a relatively realistic but less controlled setting.

Study 1

Method

Participants

Sixty participants (9 men, 51 women, median age: 22, age range: 17–30) were recruited from the Radboud University Social Sciences participant pool and later received a five-euro gift certificate or course credit for their participation. Two participants failed to complete the entire experiment and were excluded from analysis. This left 58 participants for analysis.

Effects of contingency learning and related fields such as classical conditioning and evaluative conditioning are usually medium sized to large (see, for instance, Hofmann et al., 2010), and our pilot study (Van den Brule et al., 2013) indicated we could expect a medium to large effect in this research paradigm. Therefore, the study's sample size was based on an a priori power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), such that there was an 80% chance to detect a medium to large sized effect (η_p^2 = .15) at an alpha level of .05.

Design

The robot's gestures were manipulated as a single between-subject factor with three levels: no gestures, unpredictive gestures, and predictive gestures. The participants were counterbalanced across conditions. The robot's performance was the same in each condition and mistakes could only be predicted from gestures in the *predictive gestures* condition. The *no gestures* and *unpredictive gestures* conditions served as controls. We included the no gestures condition to account for illusory correlation effects (De Jong, Merckelbach, & Arntz, 1990), where the unpredictive gestures might be subjectively perceived as predictive, despite not being predictive objectively. Any difference between the *no gestures* and *unpredictive gestures* condition would be caused by the presence or absence of the robot's gestures.

Procedure

We developed a video task in which participants watched video clips of a robot situated behind a table, on which two cans were placed. The robot's task was to select one of these cans and indicate its choice to the participants. In a single trial, the robot would go through a *contemplation* phase and a *can indication* phase. During the contemplation phase, the robot first looked at both cans and then either remained motionless or made a gesture (scratching its forehead with its right arm). Head scratching was chosen as a simple and well-known type of naturally occurring uncertainty or hesitancy behavior. It is also a noticeable gesture in the sense that it consists of an actual, clearly visible movement. In the can indication phase, the robot's normal functioning was to point at the can it had chosen. In the case of bad performance, the robot would "accidentally" push the can off the table by hitting it with its arm instead of pointing at it. Thus, the robot's performance and gesture were clearly noticeable to the participants. The videos were created and recorded with the *WeBots for Nao* simulator package (Aldebaran Robotics, 2013) and can be found in the supplementary materials.

To test whether participants were able to learn the contingency between the robot's gestures and performance, they indicated whether they thought the robot would push the can off the table on a 5-point Likert scale (1 = definitely not, 5 = definitely will), before the can indication phase was shown. Participants then watched the outcome of the robot's action in the can indication phase and continued with the next trial.

The video task consisted of 49 trials, which were distributed as shown in Table 1. In both the unpredictive gestures and predictive gestures condition, the robot performed the gesture in seven of the 49 trials, thus $P(gesture) = 1/7$ and $P(no\ gesture) = 6/7$. In all conditions, the robot pointed at the can in forty-two trials and pushed the can in seven trials, thus the probability of the robot pushing a can in each condition was $P(push) = 7/49 = 1/7$. For unpredictive gestures, this six to one ratio was preserved for the conditional probability distributions of a push following a gesture and a push following no gesture: $P(push \mid gesture) = 1/6$, $P(push \mid no\ gesture) = 6/36 = 1/6$. In the predictive gesture condition, the marginal six of the seven gestures were followed by the robot pushing a cup, thus $P(push \mid gesture) = 6/7$. The conditional probability of a push following no gesture was altered to $P(push \mid no\ gesture) = 1/42$ to ensure the marginal probability $P(push)$ was still 1/7.[1] Thus, the only difference between the predictive gestures and unpredictive gestures condition was the increased probability of a push occurring following a gesture and a push not occurring following the absence of a gesture. In the no gestures condition, the robot never made gestures, while the probability of a push action was identical to the other conditions.

[1] $P(push) = P(push \mid gesture)P(gesture) + P(push \mid no\ gesture)P(no\ gesture)$
$= 1/7 * 6/7 + 1/42 * 6/7$
$= 1/7$

Table 1: Trial distributions[2] for the different robot behaviors in Study 1.

	Trial type			
Robot behaviors	No gesture – point	No gesture – push	Gesture – point	Gesture – push
No gestures	42	7	0	0
Unpredictive gestures	36	6	6	1
Predictive gestures	41	1	1	6

After the video task, participants completed a questionnaire containing scales to measure the remaining dependent variables. The evaluation of trustworthiness was measured with a 4-item, valence-based trustworthiness scale (Ligthart, Van den Brule, & Haselager, 2013; Van den Brule et al., 2013; Van den Brule et al., 2014), Cronbach's α = .86. We also measured perceived reliability (e.g., "The robot responds the same under the same circumstances at different times," α = .76) and perceived understandability (e.g., "It is easy to follow what the robot does," α = .83) with 3 items each, based on the Human-Computer Trust Scale (Madsen & Gregor, 2000), which was adapted to an HRI context (see Appendix A). Ratings for each scale were averaged. We also asked the participants how well they could predict the robot's behavior as a single-item question. This "contingency perception" measure served to measure the participants' self-reported contingency learning. All items in the questionnaire were rated on Likert scales ranging from 1 (complete disagreement) to 7 (complete agreement).[3]

Afterward, participants were asked to supply demographics, were debriefed, and received their gift certificate or course credit.

Data Analysis

To analyze participants' predictions about the robot's actions, we used signal detection techniques. Because participants' predicted the actions of the robot on a continuous scale, we created Receiver Operating Characteristic (ROC) curves of the participants' predictions of the robot's actions (Hanley & McNeil, 1983; Stanislaw & Todorov, 1999). The area under the ROC curve is a measure of prediction accuracy (Hanley & McNeil, 1982). Areas larger than 0.5 indicate that the prediction accuracy is above chance level, whereas an area of 1 indicates perfect classification. Worse-than-chance accuracy (areas < 0.5) may be caused by sampling error, response confusion, or response bias (Stanislaw & Todorov, 1999).

The five dependent measures—*(1)* prediction accuracy, *(2)* contingency perception, *(3)* trustworthiness, *(4)* perceived reliability, and *(5)* perceived understandability—were analyzed with one-factor analyses of variance (ANOVAs) with robot behavior (no gestures vs. unpredictive gestures vs. predictive gestures) as a between-subject factor. We expected that participants in the predictive gestures condition would evaluate the robot more positively on all measures compared to participants in both control conditions (no gestures and unpredictive gestures), and that participants in the control conditions—no gestures and unpredictive gestures—would not differ significantly from each other. Therefore, we used Helmert contrasts to first compare the predictive gestures condition to the average of the control conditions (no gestures and unpredictive gestures) and then compare the control conditions to each other.

We suspected our five dependent measures would correlate and that the prediction accuracy and/or the participants' contingency perception would affect the trust-related evaluation of the robot. To identify these

[2] Note that in each condition, the overall point to push ratio is 6 to 1. In the unpredictive gesture condition, this ratio is preserved for the no gesture and gesture trials. In contrast, in the predictive gestures condition the conditional probabilities of a mistake following a gesture versus a mistake after no gesture differ, so that the presence of gestures is informative of a mistake.

[3] As an exploratory measure, implicit positive or negative associations with the robot were measured using a Single Target Implicit Association Task (ST-IAT; Bluemke & Friese, 2008; Wigboldus, Holland, & Van Knippenberg, 2006) after the questionnaire. There was no significant effect of robot behavior on implicit associations. Therefore, this measure is not further discussed in the results of Study 1 and was not included in Study 2.

potential links between the dependent measures, we conducted a path analysis using structural equation models (SEM), computed with the *lavaan* package (v0.5-17) for *R* (v3.0.0) (Rosseel, 2012). Each model's parameter standard errors were estimated with bootstrapping (1000 samples). In accordance with Hu and Bentler (1999), we report the following fit indices that describe the overall fit of the model: *(1)* The χ^2 test and associated *p*-value (non-significant values indicate good fit), which measures the discrepancy of the model's covariance matrices to the sample covariance matrices; *(2)* The Comparative Fit Index (*CFI*); and *(3)* The Nonnormed Fit Index (*NNFI*, also known as the Tucker-Lewis Index or *TLI*). For both *(2)* and *(3)*, good fit is indicated by values above 0.97, and an acceptable fit is indicated by values above 0.95 (Schermelleh-Engel, Moosbrugger, & Müller, 2003), indicating how much variance in the sample covariance matrix is accounted for by the model's predictions compared to the independence model. Finally, we also report *(4)* the Root Mean Square Error of Approximation (*RMSEA*; good fit ≤ .05, acceptable fit ≤ .08; Schermelleh-Engel et al., 2003) with its associated *p*-value (non-significant value indicates good fit), which indicates how much the model deviates from the population covariance matrix.

Results

Contingency Learning

We observed a significant effect of robot behavior on prediction accuracy, $F(2,55) = 117.1$, $p < .001$, $\eta_p^2 = .81$. In line with our expectations, planned contrasts revealed that participants predicted the robots actions more accurately in the predictive gestures condition ($M = .86$, $SD = .07$) than in control conditions (no gestures and unpredictive gestures), $t(55) = 15.29$, $p < .001$, $r = .91$. There was no significant difference of prediction accuracy between the no gestures ($M = .45$, $SD = .11$) and unpredictive gestures conditions ($M = .46$, $SD = .10$), $t(55) = 0.29$, $p = .77$. Prediction accuracy in the predictive gestures condition was significantly larger than 0.5, which indicates that the participants predicted the action of the robot above chance level, $t(18) = 22.4$, $p < .001$, $r = .98$. In the control conditions, the prediction accuracy was significantly smaller than 0.5 in the no gestures condition, $t(19) = 2.33$, $p = .03$, and the prediction accuracy was marginally significantly smaller than 0.5 in the unpredictive gestures condition, $t(18) = 1.83$, $p = .08$.

We then turned to the perception of contingency. An ANOVA showed that the effect of robot behavior on the contingency perception of the participants (i.e., how well they thought they were able to predict the robot's actions) was significant, $F(2,55) = 48.97$, $p < .001$, $\eta_p^2 = .64$. Planned contrasts revealed that the participants who watched a robot with predictive gestures reported more contingency ($M = 5.74$, $SD = 0.87$) compared to the participants who watched a robot in the control conditions (no gestures and unpredictive gestures), $t(55) = 9.48$, $p < .001$ $r = .79$. Finally, participants in the unpredictive gestures condition reported more contingency ($M = 3.05$, $SD = 1.40$) than participants in the no gestures condition ($M = 2.00$, $SD = 1.30$), $t(55) = 2.71$, $p = .009$, $r = .34$.

Robot Evaluation

Because the measures of trustworthiness, understandability, and reliability were positively correlated (*r* between .24 and .75), we used a multivariate analysis of variance (MANOVA) as an omnibus test, with the three measures as dependent variables and robot behavior as independent variable. We observed a significant effect of robot behavior on trustworthiness, understandability, and reliability, Wilks' $\Lambda = .43$, approximate $F(2,55) = 8.50$, $p < .001$, $\eta_p^2 = .24$. Follow-up univariate ANOVAs revealed a significant effect of our manipulation for all dependent variables. Furthermore, the Helmert contrasts showed the same pattern for all three scales: The robot was rated higher on trustworthiness, understandability, and reliability in the predictive gestures condition compared to the average of the two control behaviors (no gestures and unpredictive gestures), and no significant difference in score was observed between the no gestures and unpredictive gestures condition (see Table 2 and Fig. 1).

Relationship Between Explicit Evaluation and Predictability

Using SEM, we computed a path model of the relationships between the various explicit evaluation measures and the measures of predictability of the robot (Fig. 2). The exploratory model showed that the manipulations affected the evaluation of the robot's trustworthiness, understandability, and reliability indirectly through the contingency perception measure. Moreover, the prediction accuracy

Table 2. Main effect sizes (F-test statistic), effect size (η_p^2), and Helmert contrast tests (t-test statistics) of the effect of robot behavior on the explicit evaluation measures.

Measures	$F(2, 55)$	η_p^2	Contrasts $t(55)$	
			Predictive gesture vs. controls	Average of no gesture vs. unpredictive gesture
Trustworthiness	4.98 **	.15	3.10 **	0.56
Understandability	27.28 ***	.50	7.25 ***	1.39
Reliability	10.21 ***	.27	4.31 ***	1.42

Note. **p < .01, ***p < .001

measure partially mediated the effect of the predictability manipulation on the contingency perception measure. This model, without direct effects of the manipulation on the evaluation, had good fit with the data, $\chi^2(10) = 13.6$, $p = .19$; *CFI* = 0.988; *TFI* = 0.976; *RMSEA* = 0.08, $p = .30$. The Chi-square difference test with a SEM in which the direct effects were included was marginally significant, $\chi^2(6) = 11.63$, $p = .07$, which indicated that the first, more parsimonious model differed only marginally from the more complex model.

Discussion Study 1

Study 1 supported our hypothesis that naïve participants are able to learn that a robot's gestures signal poor performance when the gestures in general are predictive of making a mistake. Similarly, as expected, the use of predictive gestures improved participants' evaluation of the robot's trustworthiness,

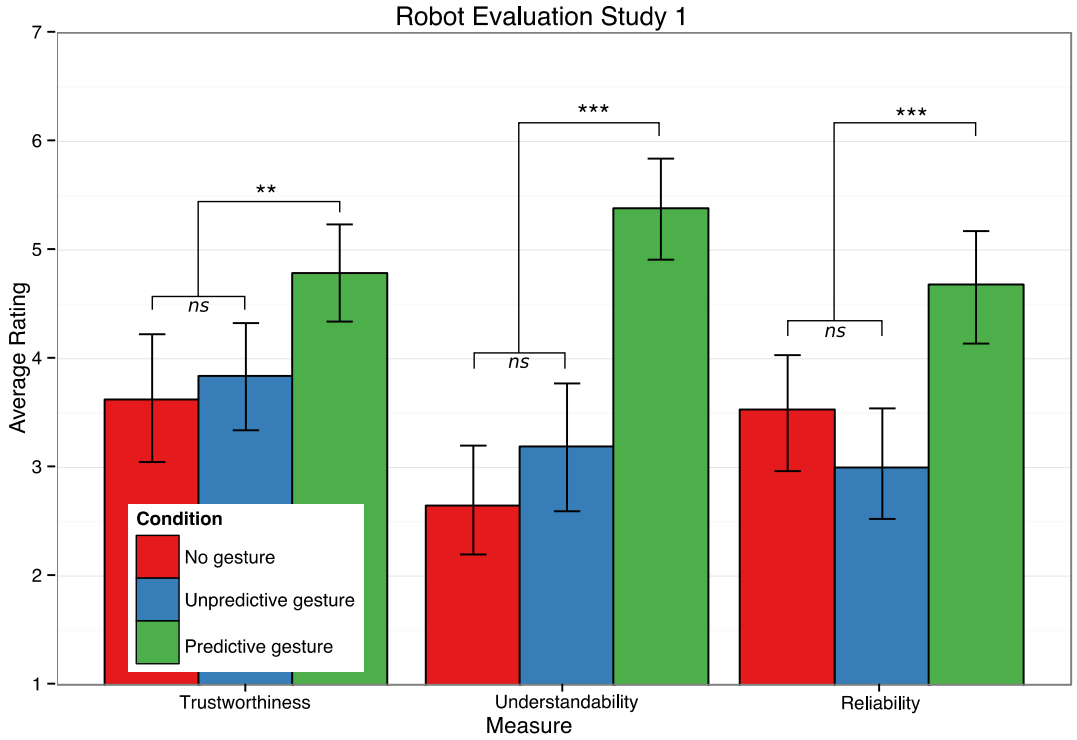

Figure 1. Means and bootstrapped 95% confidence intervals of the evaluation of trustworthiness, understandability, and reliability per condition in Study 1. The Helmert contrasts between the predictive gestures condition and controls were significant for all evaluation measures, whereas the contrasts between the no gestures and unpredictive gestures were not significant. *Note*: *ns* not significant, ** p < .01, *** p < .001.

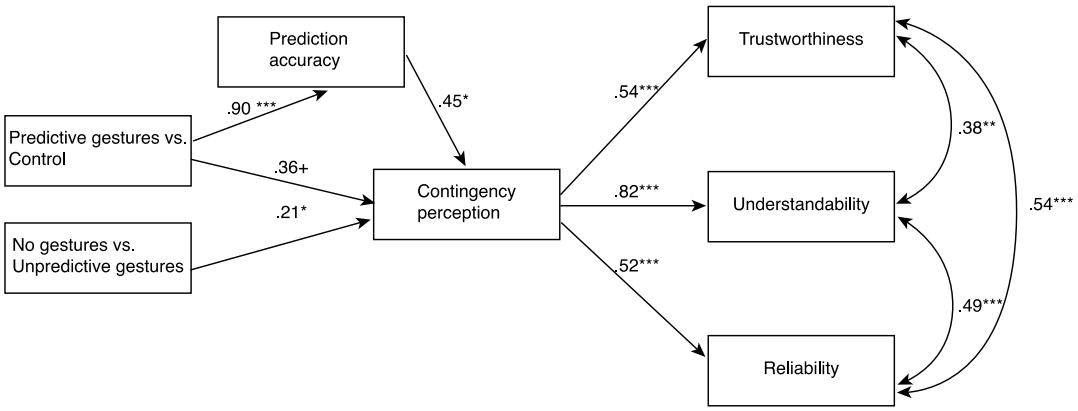

Figure 2. A path model without direct effects of the manipulations to the robot evaluation measures, showing the relationships between the various variables. Regression weights are standardized. *Note:* +$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

understandability, and reliability compared to control conditions in which the robot made no gestures or the gestures were not predictive of mistakes. No significant differences in evaluation of the robot were found between the no gestures and unpredictive gestures conditions, indicating that the presence of the gesture did not have an effect on the evaluation of the robot when it did not have a predictive value. The lack of any observed effect of the gestures themselves may indeed suggest that the meaning of the gesture may not be picked up by users when it is not coupled to mistakes. The effect of contingency is the driving force behind the pattern we observe. However, participants in the unpredictive gestures condition did believe they were able to predict the robot's actions more than the participants in the no gestures condition. This is contrary to the results of the prediction accuracy measure (which showed the participants were equally bad in predicting the robot in these two conditions) and could be caused by an illusory correlation (Chapman & Chapman, 1967; De Jong et al., 1990), a well-known psychological phenomenon in which people perceive an association between two orthogonal variables. Our exploratory path analysis suggests that the perception of contingency between behavior and performance was the critical factor. To the extent that people believe that a link between a gesture and a mistake exists, the robot is evaluated as more trustworthy, understandable, and reliable, regardless of whether the link actually exists. If the link does exist, as is the case in the predictive gesture condition, people are more likely to see a connection between gesture and mistake. Thus, equipping a robot with a system that displays gestures predictive of momentary poor performance seems to improve people's evaluation of the robot, despite the occasional poor performance.

Study 2

Study 1 showed that people are able to learn the contingency between gestures and occasional poor performance. However, learning such a contingency is especially useful for people who physically interact with a robot—especially when it enables them to improve the interaction with the robot. Furthermore, in Study 1, we explicitly asked participants to predict the robot's behavior on every trial, essentially instructing them to learn to predict. A more realistic scenario is one where people interact with a robot without being instructed to predict mistakes. Hence, in Study 2, the main goals were to test whether people would (1) learn the contingency between behavioral gesture and performance without being instructed to do so; and (2) spontaneously act to prevent or reduce the consequences of a mistake when they have learned the contingency between the robot's gestures and its mistakes. This behavior serves to improve the collaboration with the robot and should be distinguishable from other kinds of behavior of the participants during the experiment (e.g., startled reactions following a mistake by the robot, such as while building a Lego construction).

Participants engaged in an interactive version of the task with a real Nao robot. The behavior of participants was videotaped and later interpreted by raters who were blind to the experimental hypotheses. An experimenter operated the Nao robot from an adjacent room via a Wizard of Oz (WoZ)

interface. Because of the time-intensive nature of a HRI study that involves an actual robot, we only compared a robot with predictive gestures to one with unpredictive gestures in a two-condition design. Based on Study 1, we expected participants to show more preventive and goal-oriented behavior when the robot performed a gesture in the predictive gestures condition than in the unpredictive gestures condition. In contrast, we expected the participants to be less at ease in the unpredictive gestures condition because the mistakes of the robot would come as a surprise to them, as they would be unable to predict errors based on the robot's gestures. This should be reflected in more impulsive behavior of the participants compared to the predictive gestures condition.

Study 2 also served to ascertain whether the findings of Study 1 could be extended to a more realistic and interactive scenario than computer generated video clips and whether the exploratory path model found in Study 1 can be replicated. Although VHRI experiments have their merits in that they are less complex and facilitate a fast research cycle, they prevent participants from interacting with the robot in a natural way. Previous research (e.g., Xu et al., 2014) has shown that the type of media by which scenarios are presented (e.g., text, video, live interaction) can influence participants' evaluation of the robot. Moreover, ostensibly small changes in a task between modalities, such as increased interactivity in an Immersive Virtual Environment compared to VHRI, might affect the participants' perceived trustworthiness of the robot (Van den Brule et al., 2014). Thus, with Study 2, we attempted to generalize the findings of Study 1 to a more realistic setting in which the robot is an embodied entity, and the participants were able to respond and interact with the robot in less constrained manner.

Method

The method, hypotheses, and means of analysis of Study 2 were registered at the Open Science Framework before data acquisition commenced. [4]

Participants

A power calculation showed that with about 50 participants we had an 80% chance to find an effect, if it truly exists, similar in size as the smallest effects found in Study 1 (r = 0.36 and higher) in a two-group between-subject design at α = .05.

Fifty-six participants (12 men, 44 women, media age: 21, age range: 18–34) were recruited from the Radboud University Social Sciences participant pool and received a 7.50-Euro gift certificate or course credit for their participation. Participants of Study 1 were not eligible for participation. Four participants were excluded prior to analysis. Reasons for exclusion were misunderstanding the instructions, having wrong expectations about the robot's task and behavior (1 participant), and incomplete video data (3 participants). This left 52 participants for data analysis; 27 participants interacted with a robot with unpredictive gestures, and 25 participants interacted with a robot with predictive gestures.

Procedure

The task in this study was designed to be as similar as possible to Study 1. In the present study, participants interacted with a Nao selecting cups filled with Lego bricks instead of the cans used in Study 1. Participants were instructed to pick a single brick from the cup that the robot selected and to create a Lego construction with those bricks. This small change substantially increased the negative consequences of a mistake; when the robot 'accidentally' pushed a cup off the table, Lego bricks spilled across the table and onto the floor, which the participants had to pick up before the experiment continued. The interaction with the robot was recorded on video. Participants gave permission to use their recordings for subsequent analysis. Fig. 3 shows a top-down schematic of the task setup.

Participants entered the room where the experiment took place with the Nao already active in a sitting position, performing idle animations such as breathing motions and looking around. Participants read the task description from a computer screen, which contained a short video of Nao performing four trials of the task (see supplementary materials). The video was shown to let the participants get an idea of the robot's behavior during interaction and was introduced as such. Videos were shown for both robot behaviors (pointing and pushing a cup off the table). Each video showed the robot making two gestures and two

[4] Available at https://osf.io/j5dhq

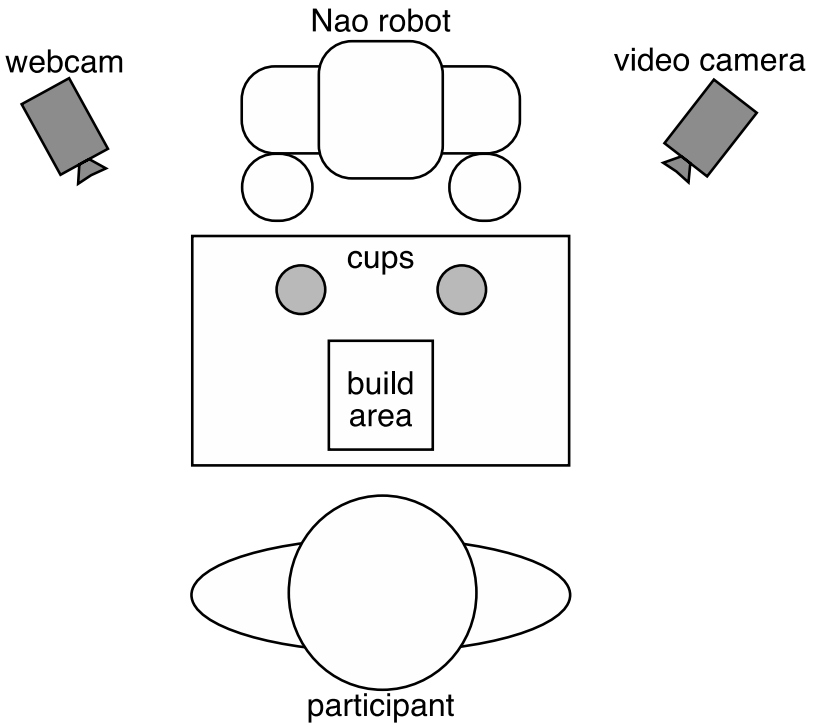

Figure 3. Top-down view of the task setup in Study 2. The participant and the Nao robot were situated opposite to each other. Two cups filled with Lego bricks stood between them. On a single trial, the robot pointed to a cup from which the participant was instructed to pick a brick to create a Lego building on the build area. The interaction was recorded with the video camera, while the webcam enabled the experimenter to monitor the progress of the experiment.

mistakes, the specifics of which depended on the condition. In the predictive gesture condition, the mistake always followed the gesture, whereas in the unpredictive gesture condition, only one of the gestures was followed by a mistake and the other mistake occurred after a gesture-free contemplation phase. The decision to show different videos depending on the condition was made so that participants would enter the next phase of the experiment already familiarized with the behavior that the robot would perform. Thus, in the unpredictive gesture condition, participants were not able to predict the outcome of the trials after watching the video, whereas the participants in the predictive gesture condition were able to get a sense of the robot's predictability by watching the videos.

Subsequently, participants were seated in front of the robot, as shown in Fig. 3. The experimenter then left the room and started the task from an adjacent room. From here, the experimenter monitored the scene in the other room through the webcam and controlled the start of the trials through a WoZ computer interface.

In order to familiarize the participants with the task, the experiment always started with two practice trials in which the robot performed no gestures and simply pointed at a cup. This was followed by a randomized sequence of trials according to the trial distributions for the condition, which are shown in Table 3. As in Study 1, the only difference between unpredictive gestures and predictive gestures conditions was the conditional probability of a push following a gesture. In the predictive gesture condition, this probability differed from the conditional probability of a push given no gesture, while in the unpredictive gestures condition these probabilities were equal. Because of time constraints, the number of trials was reduced from 49 in Study 1 to 42 in this study.

Rating of Participant Behavior

To characterize the participants' behavior during their interaction with the robot, independent raters evaluated short video clips (~11 *s* long) from the video camera recordings. Each clip started at the

Table 3. Trial distribution of Study 2.

	Trial type			
Robot behaviors	No gesture – point	No gesture – push	Gesture – point	Gesture – push
Unpredictive gestures	30	5	6	1
Predictive gestures	35	0	1	6

beginning of a trial (indicated by the moment the experimenter started the trial on the WoZ interface) and ended the moment before the robot performed its action toward a cup. The resulting clips thus contained any possible preventive, goal-oriented, and impulsive behavior of the participant in response to the robot's actions. Before the videos were rated, the videos were first screened (by the first author) to identify nonverbal movements that seemed to correspond with expectations of a mistake of the robot. Six nonverbal movements were identified: (1) reflexive movements such as startle reflexes, (2) uneasy facial expressions, (3) deliberate movements such as manipulating Lego bricks, (4) making a barrier with arms or hands to (potentially) catch falling Lego bricks, (5) arm movements to the cups, and (6) picking up the cups from the table. Based on these basic movements, three overarching behavior categories were distinguished, namely (A) preventive, (B) impulsive, and (C) goal-oriented behaviors. Preventive behaviors aimed at diminishing the consequences of the robot's anticipated mistakes and improving the collaboration with the robot. Impulsive behavior occurred when the participants were startled by the robot's mistakes. Goal-oriented behaviors are actions related to the execution of the task. Rather than requiring the raters to use a mapping between the nonverbal movements and the behavior categories, we let them rate each behavior as they saw fit based on their overall impression of the participant's action in the movie clip.

The videos of all gesture trials (7 per participant) and a random subset of the no gesture trials (12 out of 35 per participant) were rated by two research assistants who were blind to the experimental hypotheses. To ensure the assistants' judgments were not influenced by the robot's gestures, a black mask was superimposed on the robot's position and sound in the videos was muted. The assistants were unaware of the experimental manipulation.

Both assistants indicated how strongly each nonverbal movement was present in all clips. They then rated how much the behavior in each clip was impulsive and/or preventive and estimated whether the actions of the participant in the clip would potentially prevent or mitigate the consequences of a mistake. This led to nine ratings (six nonverbal movements, three types of behaviors) per clip, all of which were made on 7-point Likert scales, (1: no noticeable movement, 7: very clear movement) except for picking up cups, which was rated on a two-point scale (1: no, 2: yes). Inter-rater reliability was measured with intra-class correlations (ICC) for random observers and cases, which reached acceptable to good agreement (ICC > = .7) on five of the nine movements and behavior categories. A factor analysis with varimax rotation was done on the normalized ratings to verify the different behavioral categories.[5][6] This resulted in four factors with eigenvalues above 1, which together explained 76% of the total variance. Variables loading (Table 4) on the first dimension, which explained 27% of the variance, suggest this dimension represents behavior to prevent the negative consequences of robot mistakes. The reflexive motions and behavior all seem to load on the second dimension, which explained 22% of the total variance. The third dimension seems to represent goal-oriented movements unrelated to prevention behavior, such as picking up a Lego brick (15% explained variance), and the fourth dimension seems to represent behavior to clean up the consequences of robot mistakes (12% explained variance). The factor scores were subsequently used in the analysis of the participant behavior.

[5] Some ratings used in the factor analysis had low ICC.

[6] The reported analyses based on the resulting factor scores do not differ from analysis of the variable at interest (i.e., preventive behavior).

Results

Rated Participant Behavior

Scores of each video on of the first four factors were aggregated by trial type for each participant. Each factor was analyzed by means of a two (condition: unpredictive vs. predictive) by two (trial type: gesture vs. no gesture) mixed analysis of variance (ANOVA) with condition as between-subject factor and trial type as within-subject factor. As described below in more detail, significant effects were observed for Factors 1 and 2, whereas no significant effects were found for Factors 3 and 4, all Fs < 2.75, all ps > .10.

We expected that participants in the predictive gesture condition would perform more preventive actions when the robot performed a gesture compared to participants in the unpredictive gestures condition. More specifically, this should lead to higher scores on Factor 1 for participants in the predictive gesture condition compared to the unpredictive gesture condition. Our analysis revealed significant effects of condition, $F(1,50) = 8.26$, $p = .005$, $\eta_p^2 = .09$, trial type, $F(1,50) = 13.26$, $p < .001$, $\eta_p^2 = .10$, and, critically, a significant condition x trial type interaction, $F(1,50) = 10.34$, $p = .002$, $\eta_p^2 = .08$. Results of the post hoc *t*-tests, reported in Fig. 4, showed that participants' behavior in the predictive gesture condition during trials in which the robot performed the gesture was rated higher on Factor 1 than in the other conditions. Thus, the participants in the predictive gesture condition were able to act more preventively when the robot performed the gesture compared to the participants in the unpredictive gesture condition.

The ANOVA on Factor 2, on which ratings of reflexive movements and impulsive behavior loaded highly, showed that participants in the unpredictive gesture condition displayed more impulsive behaviors when the robot made a gesture than participants in the predictive gesture condition. The analysis revealed a main effect of condition, $F(1,50) = 4.77$, $p = .03$, $\eta_p^2 = .04$, and a marginally significant main effect of trial type, $F(1,50) = 3.55$ $p = .07$, $\eta_p^2 = .03$. The condition x trial type interaction effect was significant, $F(1,50) = 4.46$, $p = .04$, $\eta_p^2 = .04$, indicating that participants' behavior in the unpredictive gestures condition during trials in which the robot performed the gesture scored higher on Factor 2 than in the other conditions. Thisindicates that participants in the unpredictive gesture condition reacted more impulsively to the robot's gestures compared to the participants in the predictive gestures condition.

Contingency Learning

We hypothesized that if participants in the experiment would be able to learn through contingency learning that gestures of the robot were predictive of mistakes, they would employ this knowledge by displaying preventive behavior to avert serious consequences of a robot mistake. Therefore, an ROC curve analysis of the prevention behavior of the participants should lead to a higher prediction accuracy of the mistakes of the robot in the predictive gestures condition compared to the unpredictive gestures condition. Factor 1 represents behavior to prevent the negative consequences of robot mistakes. We believe

Table 4: Factor loadings of the first four factors of the factor analysis with varimax rotation. Loadings below 0.20 are not shown in this table.

Scored action unit or behavior	Factor 1	Factor 2	Factor 3	Factor 4
Action units				
Reflexive movements		0.95		
Uneasy facial expression		0.40		
Deliberate movements	0.41	0.22	0.33	
Making a barrier	0.98			
Moving arms to cup(s)			0.95	0.23
Picking up cup(s)				0.97
Behaviors				
Impulsive behavior		0.92		
Goal-oriented behavior	0.64		0.48	
Preventive behavior	0.96			

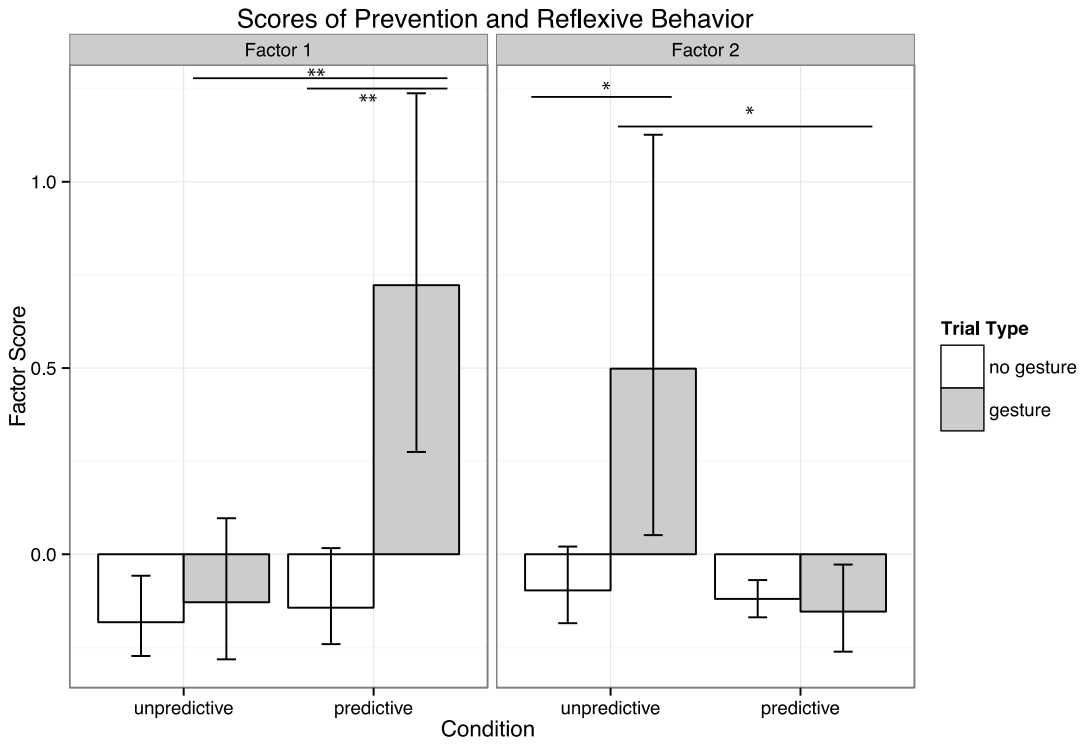

Figure 4. Mean factor scores of Factors 1 and 2 on the different trial types (no cue, cue) in each condition. Error bars represent between-subject 95% confidence intervals. *Note:* * $p < .05$, ** $p < .01$.

this factor is comparable to the predictability factor used in Study 1. In line with Study 1, we found that the prediction accuracy (i.e., area under the ROC curve) of Factor 1 was a good predictor for the participants' perception of contingency ($r = .45$ in the *SEM*). The prediction accuracy of Factor 1 was marginally larger in the predictive gestures condition ($M = 0.61$, $SD = 0.18$) compared to the unpredictive gestures condition ($M = 0.52$, $SD = 0.21$), $t(50) = 1.65$, $p = .053$, $r = .23$. Additionally, the prediction accuracy in the predictive gestures condition was significantly larger than 0.5 (which indicates higher-than-chance detection), $t(24) = 3.19$, $p = .002$, $r = .55$, whereas it was not in the unpredictive gestures condition, $t(26) = 0.57$, $p = .29$. Moreover, as in Study 1, the self-reported contingency perception was higher for predictive gestures ($M = 5.60$, $SD = 1.53$) compared to unpredictive gestures ($M = 3.19$, $SD = 1.57$), $t(50) = 5.61$, $p < .001$, $r = .62$.

Robot Evaluation

A multivariate analysis of variance (MANOVA) of the effect of robot behavior (predictive gestures, unpredictive gestures) on the robot evaluation measures (reliability, understandability, and trustworthiness) shown was significant, Wilks' $\Lambda = 0.64$, approximate $F(1,50) = 9.05$, $p < .001$, $\eta_p^2 = .36$. Further analysis with independent *t*-tests revealed the effect of condition was significant for all three scales variables (Fig. 5). In line with Study 1, the robot behavior with predictive gestures was evaluated as more reliable ($M = 4.76$, $SD = 0.86$) than the robot behavior with unpredictive gestures ($M = 3.47$, $SD = 1.17$), $t(50) = 4.51$, $p < .001$, $r = .54$. A robot with predictive gestures was also found to be more understandable ($M = 5.04$, $SD = 1.20$) than with unpredictive gestures ($M = 3.48$, $SD = 1.10$), $t(50) = 4.90$, $p < .001$, $r = .57$, and more trustworthy ($M = 4.92$, $SD = 0.83$) than with unpredictive gestures ($M = 4.14$, $SD = 1.15$), $t(50) = 2.80$, $p = .007$, $r = .37$.

Relationship Between Robot Evaluation and Predictability

We used the same path model that we found in Study 1 to confirm its validity. As was the case in Study 1, the SEM (Fig. 6) fitted the data well: $\chi^2(6) = 5.28$, $p = .51$; *CFI* = 1.0; *TFI* = 1.0; *RMSEA* = 0.0, $p = .59$;

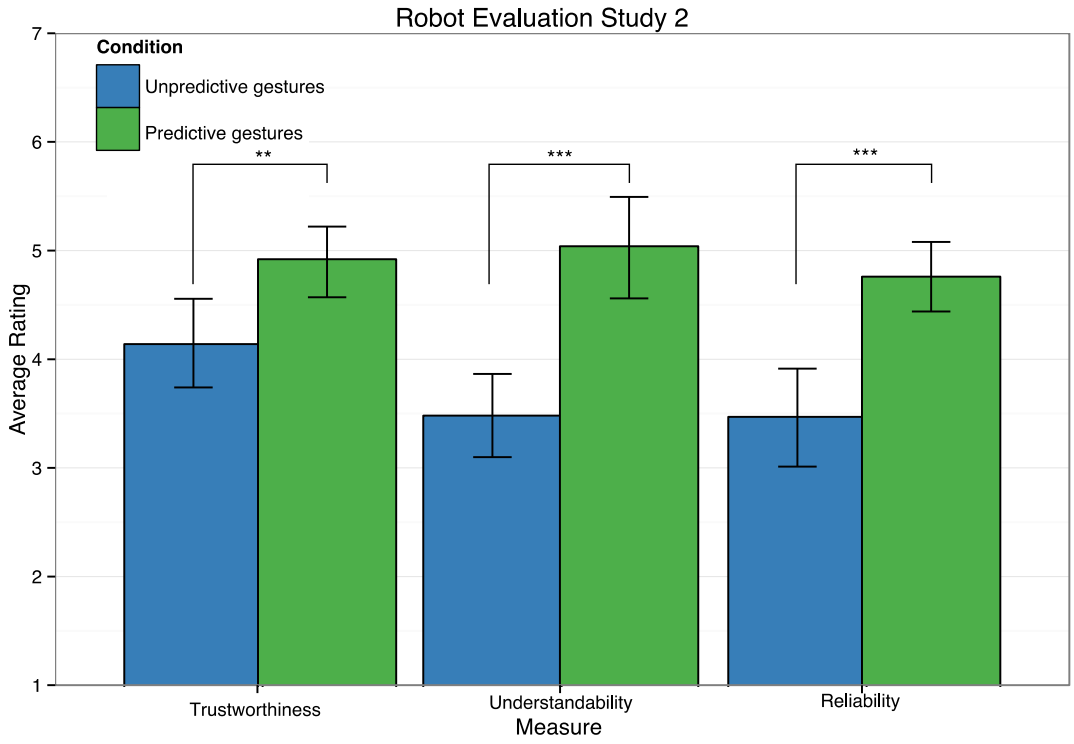

Figure 5. Means of the robot's trustworthiness, understandability, and reliability ratings in Study 2. Independent *t*-tests show a significant effect of robot behavior, all $t(50)$s > 2.797, all *p*s < .007. Error bars represent bootstrapped 95% confidence intervals. *Note.* ** $p < .01$, *** $p < .001$.

difference $\chi^2(3) = 1.05$, $p = .51$. The regression weights showed that the direct effects of the condition to the robot evaluation measures were again mediated by the contingency perception measure.

In contrast to Study 1, the prediction accuracy of prevention behavior in Study 2 did not mediate the effect of the experimental condition on the subjective perception of contingency. The resulting model thus partially confirmed the validity of the model found in Study 1.

Discussion Study 2

The results of Study 2 showed that participants learned, without any instructions to do so, that the behavioral gestures of a robot could function as warning signals when they were predictive of subsequent performance. Moreover, participants who learned that the robot's gesture is a signal for a mistake spontaneously acted to prevent or limit the consequences of a mistake made by the robot. They performed these preventive actions especially after the robot's gesture and before the actual mistake. Similar to Study 1, we found that a robot that performs predictive gestures improved the evaluation of the robot's trustworthiness, understandability, and reliability. Moreover, the extent to which participants perceived the contingency between gestures and mistakes seemed again to be the critical factor in this evaluation. Although we expected participants to be less at ease in the unpredictive gesture condition, we did not specify in advance to which types of behavior this would lead.

We did not find a correlation between the behavior of the participants during the experiment and the explicit measures from the questionnaire the participants filled in at the end of the experiment. Even though both the behavioral and questionnaire measures were affected by the experimental manipulations, the measured effects did not covary. This occurrence was also observed in our previous work (Van den Brule et al., 2014). One may conjecture that this reflects the difference between implicit and explicit knowledge (e.g., Ellis, 2008) as expressed in the participants' responses. To our knowledge, no other research in HRI exists that compares implicit with explicit measures. Psychological investigations of the implicit-explicit knowledge distinction exist (e.g., Dovidio, Kawakami, & Gaertner, 2002; Payne & Gawronski, 2010; Strack & Deutsch, 2004) and could be useful starting points for further exploration of this issue.

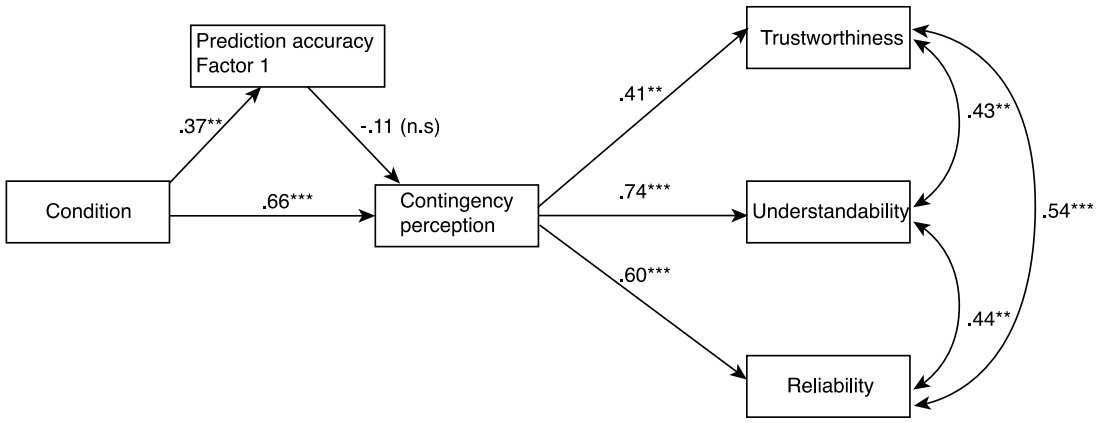

Figure 6. A path model showing the relationships between manipulations and the various outcome measures of Study 2. Regression weights are standardized. Note: *$p < .05$ **$p < .01$ ***$p < .001$.

Unexpectedly, we found that participants in the unpredictive gestures condition responded to the robot's gesture with behavior that can be characterized as impulsive. In hindsight, this may be a result from the uneasiness of the participants who were, in general, unable to predict the robot's mistakes in this condition. Previous research has indeed shown that participants are more anxious in unpredictable situations compared to predictable situations (Grillon, Baas, Lissek, Smith, & Milstein, 2004).

The video shown to the participants during the instruction phase of the experiment showed the robot acting in the way it would be acting during the interaction phase of the experiment (i.e., unpredictable in the unpredictive gestures condition, and predictable in the predictive gestures condition). This was a conscious design decision to let the participants get familiar with the behavior of the robot. However, it also makes it impossible to distinguish the effect of the video and the effect of the interaction phase of the experiment on the dependent variables collected in the study. Further work is required to disentangle these effects.

Study 2 also showed that the effects observed in the video experiment in Study 1 transferred to a more realistic setting where participants were able to interact with the robot more freely, which is in line with previous research (Woods, Walters, Koay, & Dautenhahn, 2006).

Summary and Concluding Discussion

The present research was aimed at investigating how nonverbal predictive gestures displayed by robots affect the interaction between humans and robots. Building on classical conditioning literature (e.g., Pavlov, 1927; Shanks, 2007), we predicted that people would be able to learn the signals of a robot when it makes gestures before episodes of poor performance. Furthermore, we hypothesized that this would lead to improved interaction with, and evaluation of, the robot. Studies 1 and 2 showed that naïve participants could learn the contingency between a robot's gestures and predict its performance above chance level, in Study 2 even without being instructed to do so. Study 2 showed that naïve participants spontaneously alter their behavior toward the robot to prevent or mitigate the consequences of a mistake when they learn this contingency. Participants demonstrated more preventive and goal-oriented behavior in the predictive gestures condition than in the unpredictive gestures condition in Study 2. In both studies, a robot that provided gestures (i.e., scratching one's forehead) contingently with mistakes was evaluated more positively than when no gestures (Study 1) or unpredictive gestures (Studies 1 and 2) are made. That is, robots were evaluated as more trustworthy, more understandable, and more reliable when providing predictive gestures compared to when the robot's gestures were not predictive.

In both studies, the self-reported measure of contingency and participants' behavior both showed an expected effect of robot predictability. In Study 1, we showed that participants' contingency perception mediated their robot evaluations. In Study 2, we expected that the participants' prevention behavior would show a similar mediation effect on the robot evaluation measures as the contingency perception measure in Study 1. However, in contrast with Study 1, in Study 2 there was no (partial) mediation of the prevention behavior on the relation between the condition and the self-reported contingency. The

difference in results between Study 1 and 2 could be due to the different nature of the measures; in Study 1, the participants had to explicitly predict the robot's actions, whereas in Study 2, they responded spontaneously to the robot. Therefore, the participants in Study 1 may have learned the contingency sooner, because they became aware of the experimental manipulation, whereas the participants in Study 2 were only asked about the contingency after the interaction with the robot.

The finding that participants were able to learn the contingency spontaneously, and that no deliberative process seems necessary, illustrates the potential of using contingency learning to improve HRI. Interestingly, we found no evidence that the objective prediction accuracy in Study 1, as well as the amount of prevention behavior in Study 2, affects participants' evaluation of the robot directly; only their reported contingency perception measure does. This suggests that participants evaluate the robot according to their perceived ability to predict the robot's actions regardless of their objective ability to predict the same. If this is indeed the case, it should be possible to improve a robot's evaluation by directly manipulating someone's idea that they can predict the robot's actions. For instance, people could be led to believe that the correlation between a robot's cue and making a mistake is higher than it is in reality, which is known as 'illusory correlation' (Chapman & Chapman, 1967). Even though the actual interaction might not improve by merely creating the illusion of predictability, it might lead participants to feel more at ease when they believe they can predict a social robot better than they actually can. Indeed, we did find in our experiments that to the extent people perceive that a link between a gesture and a mistake exists, they evaluate the robot as more trustworthy, understandable, and reliable, regardless of whether the link actually exists.

It may sound somewhat paradoxical that a robot's trustworthiness improves when it signals that it cannot be trusted. However, the present results demonstrate that a predictable robot is preferred despite its occasional poor performance, as long as it generally signals that a mistake might occur. This increases the transparency of the robot, which has been shown before to improve readability and evaluation of a robot system (e.g., Takayama, Dooley, & Yu, 2011). This transparency may help reducing overreliance (Parasuraman & Riley, 1997) on the system and calibrates the reliance on the robot (Lee & See, 2004, see also Van den Brule et al., 2014) and enables humans to anticipate what the robot will do, so they can intervene at the appropriate moment.

The current research has shown that it is possible to conduct HRI studies by using two different paradigms that complement one another. A video study makes it possible to quickly test an idea, while a more time-consuming interactive study confirms the effects can be observed in a more realistic setting. Some limitations of this work should also be noted. First of all, the demographics of the participant pool we had access to may not generalize well to another population, because the participants were mainly young females with a university background. Future research should investigate whether the present findings generalize to populations other than the current rather homogeneous sample of Social Sciences participants. Another point to mention is that the task that was chosen for this research may not directly resemble a real world application. The task chosen served the experimental setup very well in that it was possible to reuse its main aspects in both a video study (Study 1) and a live interaction study (Study 2), but it is unlikely that HRI will take place with the exact scenario we devised. However, the general interactive nature of the task employed may overlap with HRI situations in which division of labor and turn-taking figure prominently. However, the interactive nature of the task employed may overlap with future HRI applications. Finally, quantitative analysis of metrics derived from spontaneous behavior (as done in Study 2) is complex. In order to restrict the *'researcher's degrees of freedom'* (Simmons, Nelson, & Simonsohn, 2011), we preregistered our hypotheses and strategy for analysis. However, it turned out to be impossible to capture all the final decisions made for the analysis that was ultimately carried out, such as creating a coding scheme after the recordings took place and performing a factor analysis on the rated behaviors. Despite this limitation, we believe that the current results provide novel insights in the use of contingency learning in HRI.

In future work, it would be interesting to see what effects can be found when different signals are used. In these studies, nonverbal signals were used to communicate the robot's uncertainty. The use of nonverbal signals utilizes the implicit channel to communicate the robot's internal state, in contrast to more explicit signals such as spoken language or audiovisual cues such as flashing light and sirens. Although different signaling strategies can be used, the underlying mechanism of contingency learning is expected to be the same. Therefore, we believe that the pattern we have found here will be reproducible whenever a cue is readily discernable from a robot's normal mode of operations.

An open question remains whether the predictive gestures of the robot indeed lead to improved overall performance. In this work, we measured the influence of the robot's behavior on human evaluation and preventive behaviors. It may be interesting explore is whether the robot's predictive gestures also affect the collaborative performance on the task itself. Although the increase of human preventive behaviors suggests this is the case, a more direct measure of task performance could shed more light on this issue.

In conclusion, this work demonstrates the effectiveness of making a robot system more transparent by signaling its expected performance of a task. People learn the contingency between these signals and (potential) low performance of the robot. Predictable robots are evaluated more positively than unpredictable robots, and people are able to anticipate poor performance and take preventive actions when robots behave predictably. These findings can help in the design of domestic social robots that can be understood and trusted more by their human users.

Acknowledgements

We thank Bas Bootsma for running the Wizard of Oz experiment, and Anouk Visser and Rosanne Fikke for rating the behavioral data.

References

Aldebaran Robotics (2013). WeBots for Nao. Retrieved from http://doc.aldebaran.com/1-14/software/webots/webots_index.html

Aron, A., Aron, E., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612. Retrieved from http://psycnet.apa.org/journals/psp/63/4/596/

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71–81. doi:10.1007/s12369-008-0001-3

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*, 977–997. doi:10.1002/ejsp.487

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*(3), 193–204. doi:10.1037/h0024670

Dautenhahn, K. (2007). Methodology and themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems*, *4*(1), 103–108.

De Jong, P. J., Merckelbach, H., & Arntz, A. (1990). Illusory correlation, on-line probability estimates, and electrodermal responding in a (quasi)-conditioning paradigm. *Biological Psychology*, *31*, 201–212.

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., & Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, *20*(10), 1–8. doi:10.1177/0956797612448793

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicity prejudice and interractial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62–68.

Ellis, N. (2008). Implicit and explicit knowledge about language. In J. Cenoz & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 6, pp. 1878–1890). US: Springer. doi:10.1007/978-0-387-30424-3_143

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *54*(3), 414-420.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17695343

Feldman, R. S., & Rimé, B. (1991). *Fundamentals of nonverbal behavior*. Cambridge, MA: Cambridge University Press.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3), 143–166.

Grillon, C., Baas, J. P., Lissek, S., Smith, K., & Milstein, J. (2004). Anxious responses to predictable and unpredictable aversive events. *Behavioral Neuroscience*, *118*(5), 916–924. doi:10.1037/0735-7044.118.5.916

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. doi:10.1177/0018720811417254

Hanley, J., & McNeil, B. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*(3), 839–843. Retrieved from http://www.medicine.mcgill.ca/epidemiology/hanley/Reprints/Method_of_Comparing_1983.pdf

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Meaning+and+Use+of+the+Area+under+a+Receiver+Operating+Characteristic+(ROC)+Curve#2

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

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

Jung, M. F., Sirkin, D., Gür, T. M., & Steinert, M. (2015). Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems* (Vol. 1, pp. 2201–2210). Seoul, South Korea: Crossings. doi:10.1145/2702123.2702479

Kervyn, N., Yzerbyt, V. Y., Demoulin, S., & Judd, C. M. (2008). Competence and warmth in context: The compensatory nature of stereotypic views of national groups. *European Journal of Social Psychology*, *83*, 1175–1183. doi:10.1002/ejsp

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Ligthart, M., Van den Brule, R., & Haselager, W. F. G. (2013). Human-robot trust: Is motion fluency an effective behavioral style for regulating robot trustworthiness? In K. Hindriks, M. De Weerdt, B. Van Riemsdijk, & M. Warnier (Eds.), *Proceedings of the 25th Benelux Conference on Artificial Intelligence (BNAIC)* (pp. 112–119). Delft, The Netherlands.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In Gable, G., Viatle, M. (Eds.). *Proceedings of the 11 th Australasian Conference on Information Systems* (pp. 6–8). Brisbane, Australia.

Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

Moon, A., Panton, B., Van der Loos, H. F. M., & Croft, E. A. (2010). Using hesitation gestures for safe and ethical human-robot interaction. In *Proceedings of ICRA* (pp. 11–13). Retrieved from http://www.sites.mech.ubc.ca/~caris/Publications/Using Hesitation Gestures for Safe and Ethical Human-Robot Interaction.pdf

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. doi:10.1518/001872097778543886

Pavlov, I. P. (1927). *Conditional reflexes: An investigation of the psychological activity of the cerebral cortex*. (C. D. Green, Ed.). Oxford University Press. Retrieved from http://psychclassics.yorku.ca/Pavlov

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B. K. Payne (Eds.**).** *Handbook of implicit social cognition: Measurement, theory, and applications* (pp.1-14). New York, NY: Guilford Press.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1–5. doi:10.1037/h0025984

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Sanders, T. L., Oleson, K. E., Billings, D. R., Chen, J. Y. C., & Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1432–1436. doi:10.1177/1071181311551298

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models : Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schlencker, B. R., Helm, B., & Tedeschi, J. T. (1973). Interpersonal trust, promise credibility, and behavioral trust. *Journal of Personality and Social Psychology*, *25*, 419–427.

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology*, *60*(3), 291–309. doi:10.1080/17470210601000581

Sheridan, T. B. (1988). Trustworthiness of command and control systems. In *Proceedings of the IFAC/IFIP/IEA/IFORS Conference on Analysis* (pp. 427–431). Pergamon, Elmsford, NY.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *20*(10), 1–8.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10495845

Stanton, C., & Stevens, C. J. (2014, October). Robot pressure: The impact of robot eye gaze and lifelike bodily movements upon decision-making and trust. In *Proceedings of the International Conference on Social Robotics* (pp. 330-339). Sydney, NSW, Australia: Springer International Publishing.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *8*(3), 220–247. doi:10.1207/s15327957pspr0803_1

Sung, J., Grinter, R. E., & Christensen, H. I. (2010). Domestic robot ecology. *International Journal of Social Robotics*, *2*(4), 417–429. doi:10.1007/s12369-010-0065-8

Syrdal, D. S., Koay, K. L., Gácsi, M., Walters, M. L., & Dautenhahn, K. (2010). Video prototyping of dog-inspired non-verbal affective communication for an appearance constrained robot. In *Proceedings of the 19th IEEE International Symposium on Robot and Human Interative Communication* (pp. 632–637). Principe de Piemonte, Italy.

Takayama, L., Dooley, D., & Ju, W. (2011). Expressing thought: Improving robot readability with animation principles. In *Proceedings of Human-Robot Interaction Conference: HRI 2011* (pp. 69–76). Lausanne, Switzerland.

Tellex, S., Knepper, R., Li, A., Howard, T., Rus, D., & Roy, N. (2014). Asking for help using inverse semantics. *Robotics: Science and Systems*, *2*, 3. Retrieved from http://cs.brown.edu/courses/csci2951-k/papers/tellex14.pdf

Torrey, C., Fussell, S. R., & Kiesler, S. (2013). How a robot should give advice. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 275–282). doi:10.1109/HRI.2013.6483599

Van den Brule, R., Bijlstra, G., Dotsch, R., Wigboldus, D. H. J., & Haselager, W. F. G. (2013). Signaling robot trustworthiness: Effects of behavioral cues as warnings. In G. Herrmann, M. Pearson, A.

Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *LNCS 8239: Social Robotics* (pp. 583–584). Springer. doi:10.1007/978-3-319-02675-6

Van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., & Haselager, W. F. G. (2014). Do robot performance and behavioral style affect human trust? A multi-method approach. *International Journal of Social Robotics*, *6*(4), 519–531. doi:10.1007/s12369-014-0231-5

Wigboldus, D. H. J., Holland, R. W., & Van Knippenberg, A. (2006). *Single target implicit associations*. Unpublished manuscript.

Woods, S. N., Walters, M. L., Koay, K. L., & Dautenhahn, K. (2006). Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach. In *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control (AMC'06)* (pp. 750–755). Istanbul, Turkey: New York, NY: IEEE Press.

Xu, Q., Ng, J., Tan, O., Huang, Z., Tay, B., & Park, T. (2014). Methodological issues in scenario-based evaluation of human–robot interaction. *International Journal of Social Robotics*. doi:10.1007/s12369-014-0248-9

Yamaji, Y., Miyake, T., Yoshiike, Y., de Silva, P. R. S., & Okada, M. (2011). STB: Child-dependent sociable trash box. *International Journal of Social Robotics*, *3*(4), 359–370. doi:10.1007/s12369-011-0114-y

---

Authors' contact information: Rik van den Brule, Radboud University, r.vandenbrule@donders.ru.nl; Gijsbert Bijlstra, Radboud University**,** g.bijlstra@bsi.ru.nl; Ron Dotsch, Utrecht University, r.dotsch@uu.nl; Pim Haselager, Radboud University, w.haselager@donders.ru.nl; Daniel H. J. Wigboldus, Radboud University, d.wigboldus@socsci.ru.nl

Appendix A: Questionnaire Items

The questionnaire items below were used to measure perceived reliability and perceived understandability (both adapted from Madsen & Gregor, 2000) and valence-based trustworthiness. All items were rated on 7-point Likert scales (1 = completely disagree, 7 = completely agree).

Perceived Reliability

- The robot performs reliably
- The robot analyzes the situation consistently
- The robot behaves the same way under the same conditions at different times

Perceived Understandability

- I know what the robot will do because I understand how the robot behaves
- It is easy to follow what the robot does
- Although I may not know exactly how the robot works, I can predict whether it will pick up or knock over a cup

Trustworthiness

- How trustworthy did the robot appear to you?
- Would you entrust the robot with the task of sorting glasses?
- How positively would you judge this robot?
- How negatively would you judge this robot?