Evaluating Supportive and Instructive Robot Roles in Human-Robot Interaction

Manuel Giuliani[1] and Alois Knoll[2]

[1] fortiss GmbH, Guerickestraße 25, 80805 München
giuliani@fortiss.org
[2] Technische Universität München, Robotics and Embedded Systems,
Boltzmannstraße 3, 85748 Garching
knoll@in.tum.de

Abstract. Humans take different roles when they work together on a common task. But how do humans react to different roles of a robot in a human-robot interaction scenario? In this publication, we present a user evaluation, in which naïve participants work together with a robot on a common construction task. The robot is able to take different roles in the interaction: one group of the experiment participants worked with the robot in the *instructive role*, in which the robot first instructs the user how to proceed with the construction and then supports the user by handing over building pieces. The other group of participants used the robot in its *supportive role*, in which the robot hands over assembly pieces to the human that fit to the current progress of the assembly plan and only gives instructions when necessary. The results of the experiment show that the users do not prefer one of the two roles of the robot, but take the counterpart to the robot's role and adjust their own behaviour according to the robot's actions. This is revealed by the objective data that we collected as well as by the subjective answers of the experiment participants to a user questionnaire. The data suggests that the most influential factors for user satisfaction are the number of times the users picked up a building piece without getting an explicit instruction by the robot and the number of utterances the users made themselves. While the number of pickup actions had a positive or negative influence, depending on the role the users took, the number of own utterances always had a strong negative influence on the user's satisfaction.

1 Introduction and Related Work

When humans work together, they take different roles in the interaction. For example when two persons assemble a shelf, usually one of them takes the lead and gives instructions on how to follow the assembly plan. The other person helps the instructor to build the shelf and to gather the right parts for the next building step. The question that we are following in this publication is: in a similar construction task, do humans prefer a robot that takes the role of an instructor or a supporter?

For that, we conduct a human-robot interaction experiment in which naïve participants have to build target objects from a wooden toy construction set

B. Mutlu et al. (Eds.): ICSR 2011, LNAI 7072, pp. 193–203, 2011.

together with a robot. The robot takes either the role of an instructor or a supporter. We use the robot in both settings for a between participants experiment, in which we collect objective and subjective measurements to compare the changes in behaviour and opinion about the robot between the two experiment participant groups.

In robotics research, there are two areas in which the role of a robot is of importance: on the one hand, there are robots that have to interact with humans in various scenarios. Here, the research focuses on the different roles the robot can take in the interaction and how the human partners of the robot react to these roles. On the other hand, researchers are interested in the roles of robots in multi-robot teams. Here, robots use different roles to solve a given task more effectively.

[1] were among the first authors who realised that for a social robot that is capable to interact with a human it is of importance, which social role the robot should take in this interaction. [4] presented one of the earliest studies that researched how humans react to different robot roles. They conducted an experiment, in which a human and a robot had to work together. In the experiment, the authors varied the appearance of the robot as well as the behaviour (i.e. the role) of the robot. The results of the experiment show that humans rely more on human-like robots and feel more responsible for the task when the robot looks more machine-like. The experiment also showed that the participants felt less responsible for the task when they worked with a robot who took the role of a supervisor, which is also supported by our findings.

[9] show a socially assistive therapist robot that monitors and encourages humans in rehabilitation exercises. This robot shows either an introverted or an extroverted personality. Tapus and Maraic were able to show in an experiment that introverted patients interacted significantly longer with the introverted robot, while extroverted patients interacted longer with the extroverted robot, respectively. In contrast to our findings, it seems that in this type of interaction, humans prefer to have a partner with similar personality traits to their own.

[7] argument that for urban search & rescue robots (USAR) affective computing is important. Amongst other things, they present the theoretical basis for the implementation of social roles on a USAR, so that the robot can adapt its own behaviour on a rescue mission, corresponding to whether it is interacting with a fellow helper or a victim.

2 Human-Robot Interaction System

The experiment described in this paper makes use of a completely autonomous human-robot interaction system (**Figure 1**) which supports multimodal human-robot collaboration on a joint construction task. The participant and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech and gestures. The robot can pick up and move objects in the workspace and perform simple assembly tasks. In the scenario considered here, human and robot both know the assembly plan and jointly execute it. The robot assists the humans by explaining necessary assembly steps when the humans do not execute them by themselves and by offering pieces

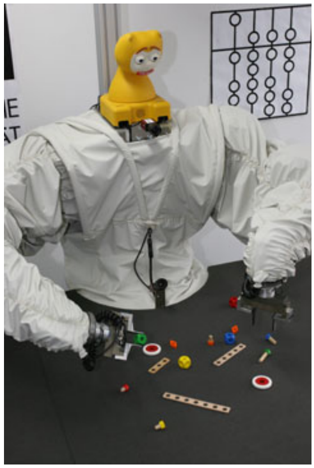

Fig. 1. Human-robot interaction system. The robot has a pair of manipulator arms with grippers, mounted in position to resemble human arms, and an animatronic talking head [10] capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech.

as required. The workspace is divided into two areas—one belonging to the robot and one to the human—to make joint action necessary for task success.

In this experiment, the robot shows two different roles: in the *instructive role*, the robot first gives instructions to the user how to assemble pieces according to the assembly plan before handing over construction pieces from its own work area. In the *supportive role*, the robot first hands over construction pieces from its own workspace to the human and only gives instructions if the users do not pick up the right construction pieces from their workspace.

Although the construction pieces can be screwed and stuck together, the robot is not able to perform assembly actions itself. However, the robot supports its human co-worker with the following list of actions: *give*, the robot hands over a construction piece from its own workspace to the human. *tellAbout*, the robot instructs the human to pick up a piece from the human's workspace because it fits to a building step of the currently loaded plan. *askFor*, the robot asks the human to put a certain construction piece on the workspace, because it is needed for a given plan and the robot cannot detect it with its object recognition. *tellBuild*, the robot asks the human to build one of the substeps of the currently loaded plan. *thankFor*, the robot thanks the human for a construction piece that the human has put on the table.

3 Experiment

In this section, we describe the experiment set-up, demography of the experiment participants, data collection and analysis, and results.

3.1 Experiment Design

This study used a between participants design with one independent variable: each participant interacted either with the robot that used the supportive role setting, or else with a system that used the instructive role. The robot was completely autonomous and did not get any other outside information than from its speech and object recognition sensors. Each participant built two target objects in collaboration with the system, always in the same order, first the *windmill* (**Figure 2a**), after that the *railway signal* (**Figure 2b**). For both target objects, the user was given an assembly plan on paper.

(a) Windmill **(b)** Railway signal **(c)** Tower **(d)** L shape

Fig. 2. Target objects for the experiment. Both target objects consist of a base that is named *tower*. A windmill is a tower combined with two small slats; a railway signal is a tower combined with an *l shape*.

The participants stood in front of the table facing the robot, equipped with a headset microphone for speech recognition. Participants got instructions that they could speak with the robot by using a set of predefined phrases: they could either ask the robot for one of the pieces in the robot's workspace by giving a direct order, for example by saying "give me a blue cube", or they could ask the robot to repeat its last utterance by saying "pardon me?".

The pieces required for the target object were placed on the table, using the same layout for every participant. The layout was chosen to ensure that there would be enough similar construction pieces on both sides of the table for every subplan of the target objects so that the robot could either perform the action *give* and handover an object from its side of the table or the action *tellAbout* and instruct the users to pick up an object from their side of the table. For example, for the tower of the windmill there was a red cube in both table areas, so that the robot could either hand over the cube from its own workspace or instruct the participants to pick up the cube from their workspace. Along with the assembly plan mentioned above, the participants were given a table with the names of the pieces they could build the objects with.

3.2 Participants

40 participants (27 male), who were naïve in the sense that they never worked with the robot before, took part in this experiment. The mean age of the participants was 27.20 (7.06), with a minimum of 17 and a maximum of 59. Of the

participants who indicated an area of study, the two most common areas were Mathematics (11 participants) and Informatics (8 participants). On a scale of 1 ("I do not agree at all") to 5 ("I completely agree"), participants gave a mean assessment of their knowledge of computers at 3.68 (1.00), of speech recognition systems at 1.90 (1.03), and of human-robot interaction systems at 1.60 (1.01). For their participation in the experiment, the participants got the chance to win a voucher for an online shop.

3.3 Hypotheses

In this study, we compare how humans react to the instructive and supportive role of the robot when they build target objects with it. We are mainly interested if the experiment participants accept both roles of the robot or if there is a clear preference for one of the two roles. In particular, we have the following two hypotheses:

H1. Experiment participants who work with the robot in the supportive role, generally asses their interaction with the robot more positive.

H2. Experiment participants who work with the robot in the supportive role display a more proactive behaviour, while participants using the instructive robot will take a more passive role in the interaction.

Since we gathered a wide range of subjective and objective measures in this study, we did not make specific predictions as to which specific measure the experimental manipulations will have an effect.

3.4 Data Acquisition

At the end of a trial, the participants responded to a usability questionnaire consisting of 29 items, which fell into four main categories: *feelings of the user* (10 items), *intelligence of the robot* (7 items), *robot behaviour* (6 items), and *task success* (6 items). The items on the questionnaire were based on those used in two previous user evaluations [2] [3], but were adapted for the scenario and research questions of the current study. The questionnaire was presented using software that let the participants choose values between 1 ("I do not agree at all") and 100 ("I completely agree") with a slider.

In addition to the questionnaire, we collected a set of objective measurements from the automatically generated system log files and from annotations of the videos we took during the experiments. All in all, we had four different objective measurements:

- the *number of verbal utterances* by the participants, which is the number of times the users asked the robot for a certain construction piece or to repeat its last utterance,
- the *number of times the participants picked up a construction piece* from their side of the table, where the robot did not instruct them to pick up the object,

- the *number of instructions the robot gave to the participants*, i.e. the instructions in which the robot told the human which piece to pick up next from the workspace, and
- the *overall duration* the participants needed to build windmill and railway signal.

We took the first two measurements from the system log files; we annotated the videos of the experiment participants with Anvil [5] to collect the remaining two measurements. Not all participants agreed that we videotaped them, thus we only have video data for 32 of the 40 participants, 17 videos of participants who used the instructive robot and 15 videos of participants who used the supportive robot.

3.5 Results

In this study, we analysed the collected data in several ways. First, we compared the subjective answers of the experiment participants to the user questionnaire to find out if there are any significant differences between the answers of the group that worked with the supportive robot and the group that worked with the instructive robot. Second, we compared the objective measurements that we took from the system logs and the videos to find differences between the two groups. Third, we calculated which of the objective measurements could potentially predict the subjective answers by the experiment participants.

Subjective Measurements. We applied a Mann-Whitney test on the answers to the user questionnaire to analyse if the different robot roles had a significant effect on the ratings by the two participant groups. Generally, participants gave a positive feedback of an average 82.84 (20.26) on the questions of the *feelings of the user* category, in which they had to rate if their interaction with the robot was enjoyable. However, the participants rated the robot's intelligence with only 56.35 (26.16) points. There was no significant difference in these questions between the two groups.

We found significant differences (p-value < 0.05) in the ratings for 4 of the 29 statements of the user questionnaire, which are displayed in **Table 1**.

Table 1. Statements with significant differences between user groups of user questionnaire

Statement	**Supportive**	**Instructive**	**M-W**
I found the robot easy to use.	83.80 (12.81)	**90.80 (13.03)**	$p \approx 0.043$
I knew what I could say or do at each point in the conversation.	71.05 (30.32)	**90.10 (12.04)**	$p \approx 0.038$
It was clear what to do when the robot did not understand me.	**70.65 (21.46)**	57.33 (15.26)	$p \approx 0.034$
The robot gave too many instructions.	33.95 (28.21)	**16.71 (21.95)**	$p \approx 0.026$

Objective Measurements. We show the results of the objective measurements in **Table 2**. We computed if there is a significant difference between the two user groups, again via a Mann-Whitney test. We found a significant difference for the number of robot instructions, which is not surprising, but shows that the instructive robot gave significantly more instructions to the user. Furthermore, users who worked with the supportive robot significantly picked up more construction pieces without getting instructions from the robot to do so.

Table 2. Objective results

Measure	Instructive	Supportive	M-W
No. of user utterances	1.65 (1.69)	1.25 (1.94)	$p \approx 0.33$
No. of user actions	0.76 (0.90)	**4.80 (1.97)**	$p < 0.01$
No. of robot instructions	**10.3 (1.49)**	4.60 (2.28)	$p < 0.01$
Assembly duration (seconds)	265.86 (46.22)	258.80 (51.32)	$p \approx 0.82$

Predictive Measurements. To complete the result analysis of this study, we calculated a predictor function to compute if the objective measurements we collected in this evaluation could predict the subjective statements of the user questionnaire. Being able to predict subjective user satisfaction from more easily-measured objective properties can be very useful for developers of interactive systems: in addition to making it possible to evaluate systems based on automatically available data without the need for extensive experiments with users, such a performance function can also be used in an online, incremental manner to adapt system behaviour to avoid entering a state that is likely to reduce user satisfaction, or can be used as a reward function in a reinforcement-learning scenario [11].

To compute the predictor function, we employed a procedure similar to that used in the PARADISE evaluation framework (PAradigm for DIalogue System Evaluation) [11]. The PARADISE model uses stepwise multiple linear regression to predict subjective user satisfaction based on measures representing the performance dimensions of task success, dialogue quality, and dialogue efficiency, resulting in a predictor function of the following form:

$$\mathit{Satisfaction} = \sum_{i=1}^{n} w_i * \mathcal{N}(m_i)$$

The m_i terms represent the value of each measure, while the $\mathcal{N}$ function transforms each measure into a normal distribution using z-score normalisation. Stepwise linear regression produces coefficients (w_i) describing the relative contribution of each predictor to the user satisfaction. If a predictor does not contribute significantly, its w_i value is zero after the stepwise process. **Table 3** shows the predictor functions that we calculated using stepwise multiple linear regression.

Table 3. Calculated predictor functions using stepwise linear regression. For calculation, four objective measurements were used, which are abbreviated in the table with *Dur* (duration to build both target objects), *Pickup* (number of anticipatory pick up actions by experiment participant), *Utt* (number of utterances by experiment participant), and *Inst* (number of robot instructions).

Measure	**Function**	R^2	**Significance**	
Feelings	$324.68 + 0.77 * \mathcal{N}(\text{Dur}) + 27.35 * \mathcal{N}(\text{Pickup})$	0.27	Dur: $p \approx 0.16$	**Utt:** $p < 0.01$
	$-40.26 * \mathcal{N}(\text{Utt}) + 20.96 * \mathcal{N}(\text{Inst})$		Pickup: $p \approx 0.16$	Inst: $p \approx 0.13$
Intelligence	$405.02 + 0.58 * \mathcal{N}(\text{Dur}) - 18.70 * \mathcal{N}(\text{Utt})$	0.15	Dur: $p \approx 0.10$	**Utt:** $p < 0.05$
Behaviour	$487.33 - 10.96 * \mathcal{N}(\text{Pickup})$	0.12	Pickup: $p \approx 0.05$	
Task success	$447.74 + 0.40 * \mathcal{N}(\text{Dur}) - 17.54 * \mathcal{N}(\text{Utt})$	0.23	Dur: $p \approx 0.10$	**Utt:** $p < 0.01$

The calculated predictor functions show that all of the objective measurements influence user satisfaction in one way or the other:

- The number of user utterances has a strongly negative influence on the three categories *feelings of the user* (abbreviated with *Feelings* in table), *intelligence of the robot* (abbr. *Intelligence*), and *task success*. The duration to build both target objects had a slight positive effect in the same three categories.
- The number of anticipatory pick up actions by the user had a positive influence on category *feelings of the user* and a negative influence on category *robot behaviour*.
- The number of robot instructions had a strong positive influence on the category *feelings of the user*, but not on the other categories.

The R^2 values of this study are in the same range as the values of our previous user evaluations. However, the values are not as high as those reported in [11] and [6].

3.6 Discussion

The results of this study show an interesting correlation: we expected that the experiment participants will prefer the supportive robot over the instructive robot (see **H1**). However, the data suggests that the users accept both robot roles and simply take the counterpart in the interaction with the robot. This can be seen from the significant answers to the statements of the user questionnaire, where the users that worked with the supportive robot answered more positive to the statement "I knew what I could say or do at each point in the conversation". This indicates that the participants showed a more proactive behaviour themselves and followed the assembly plan more by themselves when the robot gave less instructions. This is in line with the work of Hinds et al. [4], who found that humans who work with a robot that takes the role of a supervisor, felt less responsible for the task.

In contrast to that, the users who worked with our instructive robot rated the statement "The robot gave too many instructions" lower than the users from the other group, which we interpret as confirmation for hypothesis **H2**: participants who worked with the instructive robot show a more passive behaviour. One of the objective measurements also supports this claim: users who worked with the supportive robot showed a proactive behaviour and executed anticipatory pick up actions significantly more often than users of the other group. These results are in line with research from cognitive psychology and cognitive neuroscience. For example [8] review a set of studies from these fields, which also prove that humans attune their actions when working together.

The results of the calculated predictor functions are not very surprising. However, it is interesting to note that the number of anticipatory pick up actions had a positive influence on the statements in the category *feelings of the user* and a negative effect on the category *robot behaviour*. The positive influence on the feelings of the user is a confirmation for hypothesis **H1**: the participants prefer to be proactive, thus a supportive robot fits better to their preferences. The negative effect of these measurements on the assessment of the robot's behaviour can be explained with robot errors during the interaction: when the robot made an error and for example gave the wrong instructions to the user or stopped working (which could happen sometimes during the experiments because of wrongly recognised construction pieces), the users had to pick up the pieces to finish building the target objects without getting instructions by the robot.

The number of user utterances also had a negative influence on the user satisfaction. This can be easily explained: in this experiment, the system was configured so that the users did not have to speak with the robot, as long as it performed well. The users only had to talk to the robot when they either did not understand the robot's utterances and had to ask for repetition or they needed to give a direct command to the robot to ask for a piece of the robot's workspace, which almost only happened when the robot made an error. Thus, the number of user utterances is a clear indicator for problems during the experiment.

4 Conclusion

The goal of this work was to research how humans react to different roles of a robot when they have to work with the robot on a common construction task. For that, we conducted an experiment in which a human and a robot together assemble target objects from a wooden toy construction set. In this experiment we programmed the robot to take different roles in the interaction. On the one hand, the robot took the role of an instructor and gave the humans instructions on how to build the target objects before helping them by handing over appropriate construction pieces. On the other hand, the robot took the role of a supporter that directly started handing over construction pieces to its human partner and only gave instructions when necessary. To our knowledge, there have been no similar experiments conducted yet to research the role of a robot in such a construction task.

We video-taped the experiment participants and analysed the automatically generated system log files to gather a set of objective measurements from the experiment. Additionally, we asked the participants to fill out a user questionnaire to get subjective measurements as well. The analysis of the gathered data showed that, in contrast to our expectations, the users did not prefer one of the two robot roles but simply took the counterpart to the role of the robot and adjusted their own behaviour to the behaviour of the robot. This was shown in one of the objective measurements as well as in the subjective ratings of the users: experiment participants picked up construction pieces significantly more often without the robot explicitly telling them when they worked with the supportive robot; additionally, users who worked with the instructive robot wanted to hear even more instructions although the robot already gave significantly more instructions to this experiment participant group. The analysis of the influence of the objective measurements on user satisfaction revealed that in the type of scenario that we presented here, users prefer to speak less, because spoken utterances were mainly used to resolve problems in the interaction.

In future work we want to research how humans perceive different roles of a robot in scenarios in which the robot interacts with more than one human. Furthermore, we plan to analyse the arm and head movements, gestures, and verbal utterances the robot can use to emphasize its own role in the interaction.

Acknowledgements. This research was supported by the European Commission through the projects JAST[1] (FP6-003747-IP) and JAMES[2] (FP7-270435-STREP). Thanks to Sören Jentzsch for help in annotating the video data.

References

1. Breazeal, C.: Social interactions in HRI: the robot view. Systems, Man, and Cybernetics, Part C: Applications and Reviews 34(2), 181–186 (2004)
2. Foster, M.E., Giuliani, M., Isard, A., Matheson, C., Oberlander, J., Knoll, A.: Evaluating description and reference strategies in a cooperative human-robot dialogue system. In: International Joint Conference on Artificial Intelligence (IJCAI 2009), Pasadena, California (July 2009)
3. Giuliani, M., Foster, M.E., Isard, A., Matheson, C., Oberlander, J., Knoll, A.: Situated reference in a hybrid human-robot interaction system. In: International Natural Language Generation Conference (INLG 2010), Dublin, Ireland (July 2010)
4. Hinds, P.J., Roberts, T.L., Jones, H.: Whose Job Is It Anyway? A Study of Human–Robot Interaction in a Collaborative Task. Human-Computer Interaction 19, 151–181 (2004)
5. Kipp, M.: Multimedia annotation, querying and analysis in ANVIL. In: Multimedia Information Extraction (2010)
6. Litman, D.J., Pan, S.: Designing and evaluating an adaptive spoken dialogue system. User Modeling and User-Adapted Interaction 12(2–3), 111–137 (2002)
7. Looije, R., Neerincx, M., Kruijff, G.J.M.: Affective Collaborative Robots for Safety & Crisis Management in the Field. In: Intelligent Human Computer Systems for Crisis Response and Management (ISCRAM 2007), Delft, Netherlands (May 2007)

[1] `http://www6.in.tum.de/Main/ResearchJast`
[2] `http://www.james-project.eu`

8. Sebanz, N., Bekkering, H., Knoblich, G.: Joint action: bodies and minds moving together. Trends in Cognitive Sciences 10(2), 70–76 (2006)
9. Tapus, A., Mataric, M.: Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In: AAAI Spring, vol. 8 (2008)
10. van Breemen, A.J.N.: iCat: Experimenting with animabotics. In: AISB 2005 Creative Robotics Symposium (2005)
11. Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with PARADISE. Natural Language Engineering 6(3–4), 363–377 (2000)