Evaluation of an Industrial Robotic Assistant in an Ecological Environment*

Baptiste Busch[1], Giuseppe Cotugno[2], Mahdi Khoramshahi[1], Grigorios Skaltsas[3], Dario Turchi[2], Leonardo Urbano[1], Mirko Wächter[4], You Zhou[4], Tamim Asfour[4], Graham Deacon[2], Duncan Russell[2], and Aude Billard[1]

***Abstract*— Social robotic assistants have been widely studied and deployed as telepresence tools or caregivers. Evaluating their design and impact on the people interacting with them is of prime importance. In this research, we evaluate the usability and impact of ARMAR-6, an industrial robotic assistant for maintenance tasks. For this evaluation, we have used a modified System Usability Scale (SUS) to assess the general usability of the robotic system and the Godspeed questionnaire series for the subjective perception of the coworker. We have also recorded the subjects' gaze fixation patterns and analyzed how they differ when working with the robot compared to a human partner.**

I. INTRODUCTION

Robotic assistants are gaining popularity and their usage has increased in the past few years, with applications ranging from human guidance [1] to personal care [2]. In industrial settings, however, the applications mainly concern the automation of repetitive and dangerous tasks, although the modalities for possible interaction are far richer [3]. Unlike their predecessor, future industrial robots are meant to work alongside and collaborate with humans on a daily basis. Evaluating their social impacts on their human coworkers is thus of prime importance.

The social evaluation of robotic systems is a broad field of Human-Robot Interaction (HRI) [4], [5], [6]. Often, the evaluation is limited to the usability of the control interface [7], or the safety of industrial coworkers [8]. However, it is crucial to assess also the acceptance of the robotic systems as it plays a key role in the interaction. Such evaluations help with the design of new effective robotic platforms in terms of hardware and behavior [9], [10]. Anthropomorphism, the tendency to give human-like traits to inert objects or living organisms, impacts our perception of various technologies [11]. Based on this, humanoid robots, as opposed to other designs such as spider-leg robots, would therefore be better accepted by their coworkers [12]. Robotic motions, especially of the head, are also of interest as they convey information to the human partner [13]. Another important factor is the coordination between partners, especially for carrying tasks such as the one depicted in Fig. 1. Related to this, haptic communication and force coupling has been linked to increased coordination [14]. Moreover, one of the factor that impacts the most the acceptance on a long term is the usability of the whole system [15].

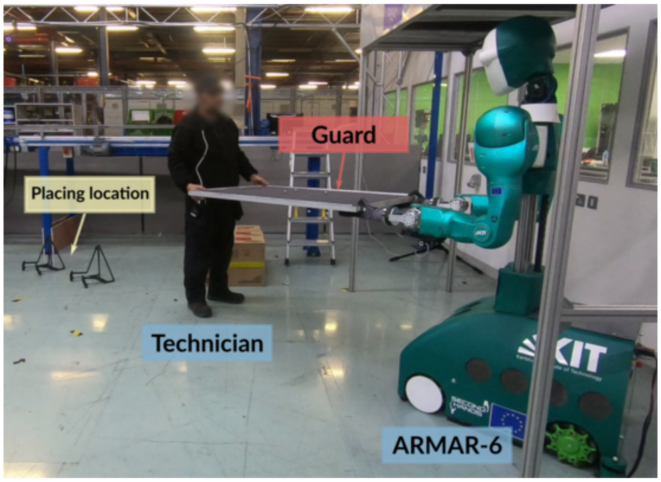


Fig. 1: ARMAR-6 providing help during the guard removal sequence. The goal is for both agents (blue) to carry the guard (red) to the expected location (yellow).

To assess for the usability of a general technology, a common metric is the System Usability Scale (SUS) [16]. This metric is, however, very generic and not particularly tailored to evaluate robotic systems. Usability of robotic systems has been studied in telepresence robotics [17], or social care robotics [18], but has been limited in industrial settings. With robotic assistants, evaluating how the robot is perceived by its partners is also an important metric. Most of the research in this area has focused on assistant robots acting as caregivers [19], [20]. Beside subjective evaluation, it is important to consider quantitative metrics such as gaze fixations. For example, horizontal gaze deviations have been commonly used to evaluate workload and trust in autonomous driving vehicles [21]. The notion of trust in the robotic systems is also correlated with the robots performance and faults which can be measured quantitatively [22].

In this research, we evaluate a fully autonomous industrial robotic assistant, ARMAR-6, developed by KIT as part of the SecondHands project[1], and presented in Fig. 1. It is a torso-humanoid platform made of two 7 DOF arms mounted on a holonomic movable base. The on-board algorithms comprise object detection, human detection, and voice recognition

*This work was supported by European Union under project Horizon 2020 SecondHands grant agreement 643950

[1]B. Busch, M. Khoramshahi, L. Urbano, and A. Billard are with EPFL Learning Algorithms and Systems Laboratory (LASA), Lausanne, Switzerland

[2]G. Cotugno, D. Turchi, G. Deacon, and D. Russell are with the Robotics Research Group of Ocado Technology, Hatfield, UK

[3]G. Skaltsas is with University of Hertfordshire, Hatfield, UK

[4]M. Wächter, Y. Zhou, and T. Asfour are with KIT High Performance Humanoid Technologies Lab (H2T), Karlsruhe, Germany

[1] `https://secondhands.eu/`

software. For full technical specifications, we refer to the original article that describes the platform [23].

Acceptance was considered as a key factor in designing hardwares, algorithms, and controllers for this robotic platform. The humanoid design is expected to ease its integration in industrial settings, in terms of both acceptance, and the capacity to handle industrial tools. For instance, natural language processing is used as a means to communicate with the human technicians. To account for force coupling and haptic communication, we use dynamical systems as a motion generator. This allow smooth switching between tasks, and ensures coordination between the robot and its coworker [24].

The goal of this research is to evaluate our robotic platform (including its algorithms/controllers) at its current stage of development in an ecological environment. As part of the SecondHands project, the robot is designed to be used by Ocado, an online supermarket, to support workers with maintenance tasks. The robot is supposed to be able to carry parts of a diverter with its human partner and provide him/her with the necessary tools. Therefore, the evaluation was carried out at Ocado facilities with the real end-users.

During the evaluation we have focused on three main components: (1) the usability of the system, (2) the perception of the robot by its coworker, and (3) the impact of the interaction on the partner's gaze fixations. For the latter, we hypothesize that coworkers should share similar fixation patterns when working with the robot as with a human partner as this could be linked to anthropomorphism. Differences between the fixation patterns could be a metric for mistrust and/or novelty effect due to working with a robot for the first time. We expect those differences to decrease with the number of repetitions of interaction.

Usability of the system was assessed using a modified version of the SUS [16], tailored to the robotic application, and perception of the robotic system with the Godspeed Questionnaire Series [25]. We have also recorded and analyzed gaze fixations specifically the fixation time, of the human coworker while they performed the same task with a human coworker or with the robotic assistant. The full experimental protocol is detailed in Section II. In Section III we detail the collected data and their analysis. Results are provided in Section IV and discussed in Section V.

II. Experimental protocol

The goal of this study is to evaluate the usability of the robot and how it is perceived by its coworker. To this extent, it comprises two surveys, a modified version of the SUS [16], tailored to the robotic application, and the Godspeed Questionnaire Series [25], to get a subjective evaluation from people performing the task. Looking at correlations between the two questionnaires is expected to shed some light on what aspects of the robot are needed to be improved for a better usability of the system. Failure events appearing during the interaction were also recorded. We hypothesize that they are linked to lower usability ratings. Finally, gaze fixation patterns, recorded with an eye-tracker system, are analyzed to see how they differ when working with the robot compared to when working with a human partner.

Six maintenance technicians(all male, aged 22-51) and seven engineers, referred as lab members(all male, aged 29-54) were recruited for the study from Ocado. Technicians were familiar with the task as it is one of their regular job requirements, although the mock-up diverter is a simplified version of the one they are normally working with (e.g. simplified way to unlock the guard).

The technicians had never interacted with the robot prior to the experiment, whereas the lab members had never previously done the maintenance task. However, some of the lab members had experience interacting with robotic systems, but never with ARMAR-6.

The experiment included the material listed below.

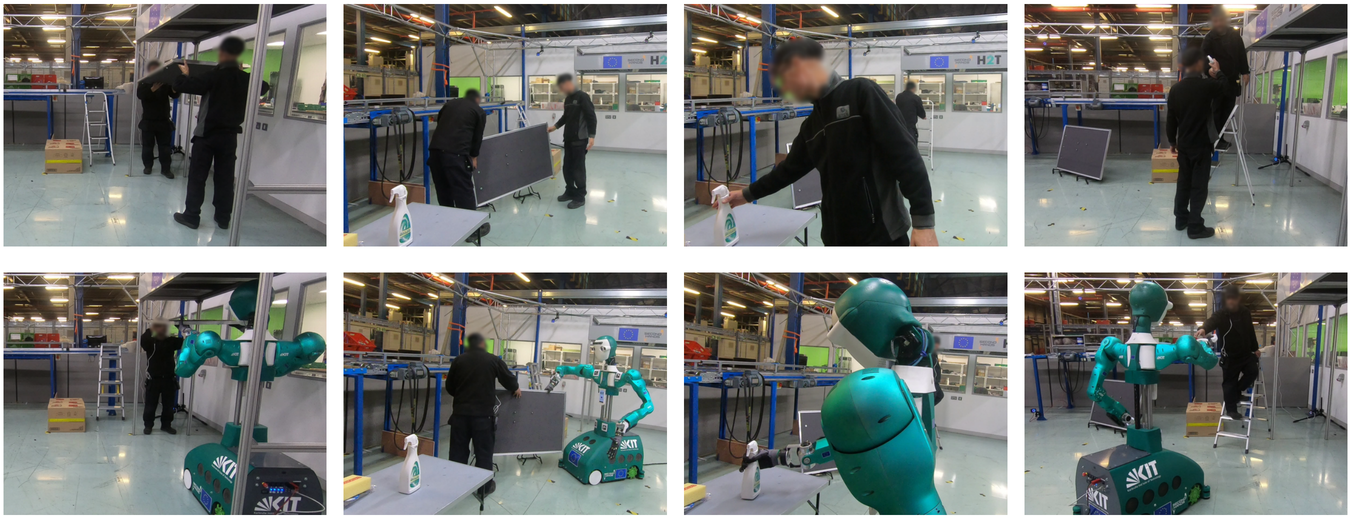
- 1 robot ARMAR-6 platform equipped with voice and human posture recognition engine and inboard cameras
- 1 eye-tracker from Pupilabs
- 1 Optitrack tracking system
- 2 external cameras (GoPro Hero 6)
- 1 microphone with headset
- 1 mock-up diverter built by KIT
- 1 spray bottle
- 1 ladder
- 1 table

The list of measurements is provided below and detailed in Section III.

- Failure events logged on paper
- Gaze pattern (from the eye-tracking device)
- Godspeed questionnaire, asked both prior the interaction and after
- Modified SUS asked only after the experiment
- Free comments

The task was divided into two parts and showcased two partners performing the maintenance of a diverter as shown in Fig. 2. The first part consisted of removing the guard from the diverter and placing it at a specified location (see Fig. 2a-2b). We refer to this sequence as the **guard removal** sequence. The second part was a cleaning task, where the technician had to climb on a ladder and use a spray bottle and cleaning cloths to clean the diverter (see Fig. 2c-2d). We refer to this sequence as the **cleaning** sequence. A single run of the experiment lasted approximately 5 minutes.

Subjects in the study were divided into pairs to perform the task, keeping the two groups of subjects separated. Each subject performed the task four times, once with his paired partner, and three time interacting with the robot. During the human-human interaction, the studied subject was assigned the role of the leader, while the other one was the follower. When one subject was performing the task with the robot, his paired partner was asked to wait in a separate room. Afterwards, the first subject played as the follower for the human-human condition of the second one. As the number of lab members was not even, one of the experimenter had to play the follower role during the human-human condition of the 7^{th} lab member. The robot always assumed



(a) lowering the guard (b) placing the guard (c) grasping the bottle (d) handing-over the bottle

Fig. 2: Pictures of the key parts of the task in both human-human (top) and human-robot (bottom) conditions with the same technician. (a-b) The guard removal sequence where both partners have to carry the guard to a desired location. (c-d) The cleaning sequence where the technician cleans the diverter while standing on a ladder.

the follower role and replied to orders/commands from the leader. Commands were given via the headset microphone and translated using our voice recognition software.

Prior to the interaction, the subjects were briefed all together with a short explanation of the setup. The briefing consisted of slides covering the following points:

- a video recording of the task, shot during a live demonstration at CEBIT 2018[2]. This video gave the subjects some insights into what to expect during the task
- a list of risks associated with the interaction
- a list of the voice commands expected by the robot during the task
- an information slide about the data collected.

After the briefing, participants had to sign a consent form, prior to start the experiment.

III. Method

We have used the two external cameras to record the interaction, as well as the eye-tracking device to capture the gaze pattern of the subject interacting with the robot. The robots internal camera was also used to record the interaction from its viewpoint. We have used the videos to check that no failure events were missed, and the eye-tracking data to analyze the gaze fixations of the leading subject.

In the following paragraphs, we provide additional details on the recording process and data analysis for each measurement performed.

A. *Failure events logging*

During the evaluation, each event that lead to a failure of the run (i.e. the robot could not recover and the run had to be restarted) were logged. Partial success, when at least one failure event appeared during the interaction but the experiment could be recovered, were also noted and distinguished from a complete success. A failure event could be either the result of an algorithmic error (e.g. voice recognition mis-recognized a sentence) or a mechanical problem (e.g the guard slips from the robot hand). These are important elements to consider to determine the principal problems to correct for, for a safe and more natural interaction.

B. *Eye-tracking video labelling*

During both human-human and human-robot interactions, the leader subject was wearing an eye-tracking device to record gaze fixations. Videos of the leader's point of view were recorded from the frontal camera of the device, and gaze fixations were reconstructed and projected onto the videos using the provided software from Pupilabs. A post processing step was performed to manually annotate the gaze fixations, using BORIS video annotation software [26]. The labels are chosen among the following categories:

- Right hand (robot or human partner)
- Left hand (robot or human partner)
- Head (robot or human partner)
- Torso (robot or human partner)
- Object (guard or spray bottle)
- Visual servoing
- Direction of motion

Visual servoing consisted of fixations were the leader was looking at his own hands, e.g. while unlocking the guard from the diverter. Direction of motion consisted of fixations were the leader was looking at where he was heading during, for example, transportation of the guard or placement at the desired location. As the time for each run is different, and a large difference appears between human-human runs

[2] https://secondhands.eu/index.php/the-secondhands-project-at-cebit-2018/

and human-robot ones, the results are normalized with the time spent at each fixations by the total time of the run. Recordings were separated between runs, i.e. we gathered data for the human-human condition and the three repetitions with the robot.

We use this metric as a means to evaluate the effects of interacting with the robot on the gaze patterns. One hypothesis we had is that interacting with the robot has an effect on the gaze patterns of the technician, due to the lowest speed execution, lack of trust, and probably the novelty effect, but this effect should decrease the more subjects interact with the robot, to converge to their normal behavior. A Wilcoxon signed-rank test was used to evaluate the paired differences between the human-human condition and each repetition.

C. *Godspeed questionnaire series*

The Godspeed Questionnaire was asked to be completed prior to the experiment, after seeing the video of the robot interacting with a technician recorded during CEBIT, and after the interaction. The questionnaire comprises 24 Likert scale items ranging from one to five, where the one and five represent opposites (e.g. Ignorant-Knowledgeable, 1 means Ignorant and 5 means Knowledgeable). There are four sets of questions that cover **Anthropomorphism**, **Animacy**, **Likability**, **Perceived intelligence**, and **Perceived safety**. The last set of questions, i.e. **Perceived safety**, are not asked in a way to rate the robot impression ("Please rate your impression of the robot on these scales") but to rate the subjects' own emotional state ("Please rate your emotional state on these scales"). Words on the left are traits that mostly characterize machines, whereas words on the right mainly characterize humans. For comparison purposes of the results we have subtracted 1 to all questions to have a final score between 0 and 4. This questionnaire is a means of evaluating the effect of the interaction on the perception of the robotic system.

Results of the surveys were divided into the two groups of subjects, **technicians** and **lab members**, as well as pre and post interaction groups. Statistical analysis has been performed (Wilcoxon signed-rank test) to highlight the paired significant differences between the results of pre and post interaction for both groups. Additionally, a Kruskal-Wallis test has been performed to highlight the non-paired differences between the results of the the two groups.

D. *Modified System Usability Scale*

The modified SUS survey was only filled out after the experiment. It comprises 10 Likert scale affirmation items, where scores range from Strongly disagree (1) to Strongly agree (5). Changes made to the original SUS scale consist mostly in replacing "the system" with "the robot". Question 4, 7 and 10 are the one that have been the most modified. In the original version, those questions insist on the learning process to use the system correctly. Because technicians already know how to perform the maintenance task, we were more interested in knowing how the robot impact their working conditions and modified the questions to focus on those aspects. As even questions are asked in a negative form, two transformations are required to compare the results:

- Substract 1 from the values of odd questions
- Substract the values of even questions from 5

After those transformations, we obtain numbers ranging from 0 to 4 where 4 is the best value to obtain for both types of questions. To obtain the final usability score, according to the SUS methodology, we need to sum all the questions and multiply the final value by 2.5. This gives a final result between 0 and 100, 100 meaning perfect usability. A Kruskal-Wallis test has then been performed to highlight the non-paired differences between results of the the two groups.

E. *Free comments*

After the interaction and filling out of the questionnaires, subjects could anonymously leave their comments on a sheet of paper. No special analyses were performed on those comments, but they provide important information to be used for future developments.

IV. Results

In the next paragraphs we detail the results gathered, for all the measurements.

A. *Failure events*

Table I shows the repartition of runs in failed, partial success and success during both the **guard removal** and **cleaning** sequences for both groups.

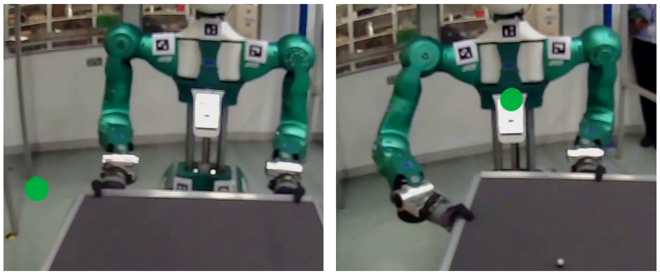
Technicians	**Failed**	**Partial**	**Success**
Guard removal	8	7	3
Cleaning	3	3	4

Lab members	**Failed**	**Partial**	**Success**
Guard removal	11	8	2
Cleaning	2	2	6

TABLE I: Number of failure events during guard and cleaning sequences

During the **guard removal** sequence, most of the failure events were caused by the "re-grasping" action, shown in Fig 3, where the robot had to rearrange the position of its hands to ensure a successful placement of the guard at the desired location, but was closing its hands in the air (41% of the failure events), or by the guard falling down when placing due to an excessive weight on the supporting hand (41% of the failure events as well).

For the **cleaning** sequence, most of the failure events were caused by a voice recognition problem (42% of the failure events), e.g. the subjects using a different sentence to trigger the action, or the robot mis-recognizing the sentence. The rest of the failure events were appearing during the handover (38% of the failure events), due to the subjects being too high on the ladder or being too tall to be fully visible.



(a) Before re-grasping

(b) After re-grasping

Fig. 3: Pictures of the re-grasping event from the eye-tracker showing the robots grasp of the guard before (left) and after (right) re-grasping.

B. Gaze fixations

Looking at the results in Fig. 4a, we observe that subjects spent more time performing visual servoing in the human-human interaction compared to the human-robot one ($p = 0.028$, Wilcoxon signed rank test). There is also differences in the fixations on the right hand ($p = 0.018$). For the repetitions of the experiment with the robot, shown in Fig. 4b, there are no significant differences between runs.

C. Godspeed questionnaire series

The statistical analysis on the difference between pre and post interaction shows the following:

- Technicians found the robot to be more lifelike (from $avg = 1.0 \pm 1.15$ to 1.66 ± 0.94, $p = 0.025$) and more responsive (from $avg = 1.5 \pm 0.76$ to 2.16 ± 0.68, $p = 0.038$)
- Lab members found the robot to be less competent (from $avg = 2.0 \pm 0.53$ to 1.28 ± 0.45, $p = 0.046$) and less intelligent (from $avg = 1.86 \pm 0.98$ to 1.0 ± 0.53, $p = 0.045$)

The rest of the differences are not significant ($p > 0.05$). Therefore, in Table II we show only the average results after interaction.

In terms of differences between the two groups, only the questions **Artificial-Lifelike**, **Inert-Interactive**, and **Quiescent-Surprised** are significant. The first difference is that technicians found the robot more lively after interaction. Looking at the results of the two other questions, we observe that all the technicians answered 3 for both (middle of the scale) which lead to no variance in their answers. Therefore, the differences on those questions might be an artifact.

D. Modified SUS

Fig. 5 shows the average results for both technicians and lab members on the modified SUS survey.

There are no significant differences between the two groups on any of the questions. The final score for technicians is $avg = 54.16 \pm 11.42$, and $avg = 47.5 \pm 14.01$ for lab members.

After analysis, we observe that they are correlations between some items of the Godspeed series and some of the SUS questions. For lab members, this mainly concerns traits that are linked to robot intelligence, e.g. the Godspeed item **Incompetent-Competent** and the SUS item **I felt that interacting with the robot was natural** are positively correlated, so is **Ignorant-Knowledgeable** and **I would like to work with the robot frequently**. This suggest that lab members that have rated the robot as more intelligent also rated it as more natural to interact and that can be used on frequently. For technicians the correlations are more linked to anthropomorphism and animacy, e.g. the Godspeed items **Machinelike-Humanlike** and **Unconscious-Conscious** are both negatively correlated with the SUS item **I thought the robot was a good coworker**, suggesting that those who rated the robot as more human-like also rated it as a bad coworker. Tables III and IV show the most correlated items for technicians and lab members respectively.

A correlation analysis with the answers to both surveys and the number of **failures**, **partial success**, and **success** events for each subject show that there exists, for technicians only, a negative correlation between the number of failures and the Godspeed item **Agitated - Calm** (correlation coefficient -0.95), suggesting that technicians who felt a bit safer were also the ones who experienced the least number of failures.

E. Free comments

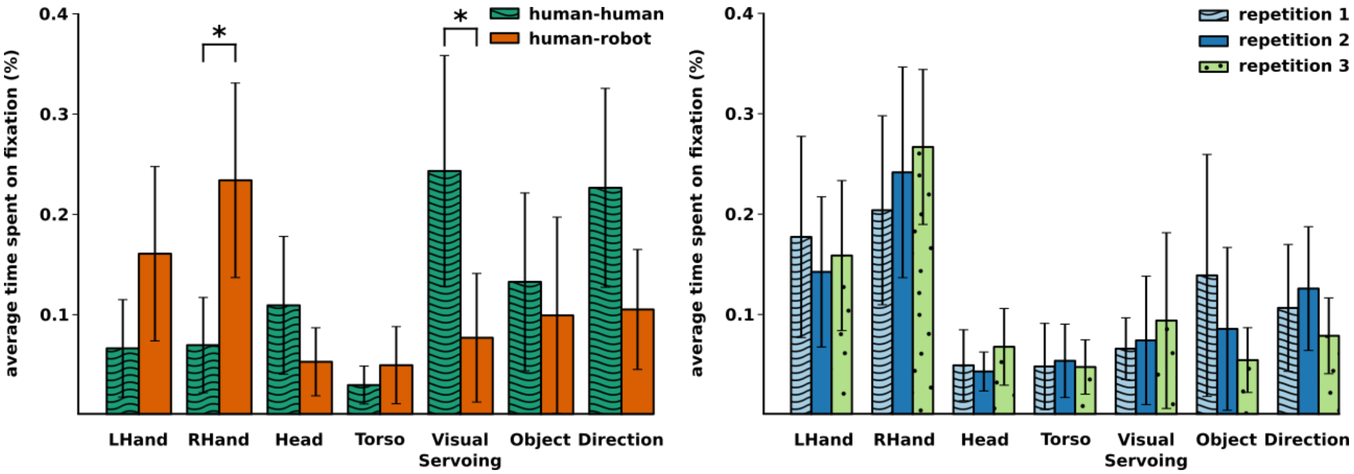
Free comments could also be left on a sheet of paper although almost only lab members left comments. Summarizing the comments reveals the following:

- The robot is responsive and interaction via forces (push/pull) felt natural with smooth motions
- The scenario of the interaction is too scripted especially during voice interaction (fixed sentences that do not account for semantics)
- The robot was acting too much as a follower and sometimes needed the help of the human (e.g. re-grasp)

V. Discussion

As suggested by the different results, improvements have to be made in terms of the usability of the system. The usability score of the robot, at its current stage ($avg = 50.83$ when averaging over both groups), is a bit small compared to the average score of products tested with the SUS scale ($avg = 68$) [27]. However, we have to note that there are currently no standards in the HRI literature to assess the usability of robotic systems and the SUS scale had to be slightly modified to target our specific needs.

The results of the Godspeed evaluation are encouraging, as subjects were rating their likability of the system above average. Comparing the results with the rest of the HRI literature is a bit challenging as ARMAR-6 platform is a mix between an industrial and a social robot. The Godspeed questionnaire has been extensively used in the assistive robotics field [5], but there is no mention of it being used to evaluate industrial robots. Nevertheless, some of the effects we have reported have also been observed in other settings. Decreases in terms of the perceived intelligence pre and post interaction have been observed in conversational robots [28], and lifelike



(a) Human-human and human-robot

(b) Human-robot repetitions

Fig. 4: Average and standard deviation of the time spent fixating (in % of the total time of the run) in both human-human and human-robot conditions (left) and in each human-robot repetitions (right). Significance has been tested using Wilcoxon signed rank test.

Anthropomorphism	Technicians	Lab Members	Animacy	Technicians	Lab Members
Fake - Natural	1.17 ± 0.4	1.28 ± 1.11	Dead - Alive	1.34 ± 0.52	1.71 ± 1.5
Machinelike - Humanlike	1.34 ± 0.82	0.86 ± 0.69	Stagnant - Lively	1.0 ± 1.1	1.71 ± 0.49
Artificial - Lifelike	1.67 ± 0.52	1.0 ± 0.81	Mechanical - Organic	1.17 ± 0.75	0.86 ± 0.69
Unconscious - Conscious	1.67 ± 1.03	1.14 ± 0.69	Artificial - Lifelike	1.67.0 ± 0.82	0.71 ± 0.49
Moving rigidly - Moving elegantly	1.67 ± 0.82	1.14 ± 1.07	Inert - Interactive	2.0 ± 0.0	1.43 ± 0.53
-	-	-	Apathetic - Responsive	2.17 ± 0.75	2.14 ± 0.9
Likeability	**Technicians**	**Lab Members**	**Perceived Intelligence**	**Technicians**	**Lab Members**
Dislike - Like	2.17 ± 0.75	2.57 ± 0.53	Incompetent - Competent	1.34 ± 0.81	1.28 ± 0.49
Unfriendly - Friendly	2.16 ± 0.75	2.14 ± 0.69	Ignorant - Knowledgeable	1.67 ± 0.81	1.28 ± 0.75
Unkind - Kind	2.34 ± 0.52	1.86 ± 1.21	Irresponsible - Responsible	2.17 ± 0.75	1.86 ± 0.69
Unpleasant - Pleasant	2.5 ± 0.55	2.28 ± 0.95	Unintelligent - Intelligent	1.67 ± 0.82	1.0 ± 0.58
Awful - Nice	2.33 ± 0.52	2.28 ± 0.75	Foolish - Sensible	2.16 ± 0.41	1.71 ± 0.95
Perceived Safety	**Technicians**	**Lab Members**	**Average**	**Technicians**	**Lab Members**
Anxious - Relaxed	1.33 ± 0.81	2.14 ± 1.06	Anthropomorphism	1.5 ± 0.72	1.09 ± 0.84
Agitated - Calm	1.83. ± 0.75	1.85 ± 1.07	Animacy	1.56 ± 0.8	1.43 ± 0.93
Quiescent - Surprised	2.0 ± 0.0	1.14 ± 0.69	Likeability	2.3 ± 0.58	2.23 ± 0.93
-	-	-	Perceived Intelligence	1.8 ± 0.75	1.42 ± 0.73
-	-	-	Perceived Safety	1.72 ± 0.65	1.71 ± 0.98

TABLE II: Results of the Godspeed questionnaire series (average and standard deviation) for both groups

Godspeed item	SUS item	Correlation coefficient
Moving rigidly - Moving elegantly	I found working with the robot unnecessarily complex	0.94
Machinelike - Humanlike	I thought the robot was a good coworker	-0.92
Unconscious - Conscious	I thought the robot was a good coworker	-0.92
Artificial - Lifelike	I would like to work with the robot frequently	-0.94

TABLE III: Correlations between Godspeed and SUS (Technicians)

Godspeed item	SUS item	Correlation coefficient
Incompetent-Competent	I felt that interacting with the robot was natural	0.88
Ignorant-Knowledgeable	I would like to work with the robot frequently	0.87
Apathetic-Responsive	I felt uncomfortable working with the robot	-0.98

TABLE IV: Correlations between Godspeed and SUS (Lab members)

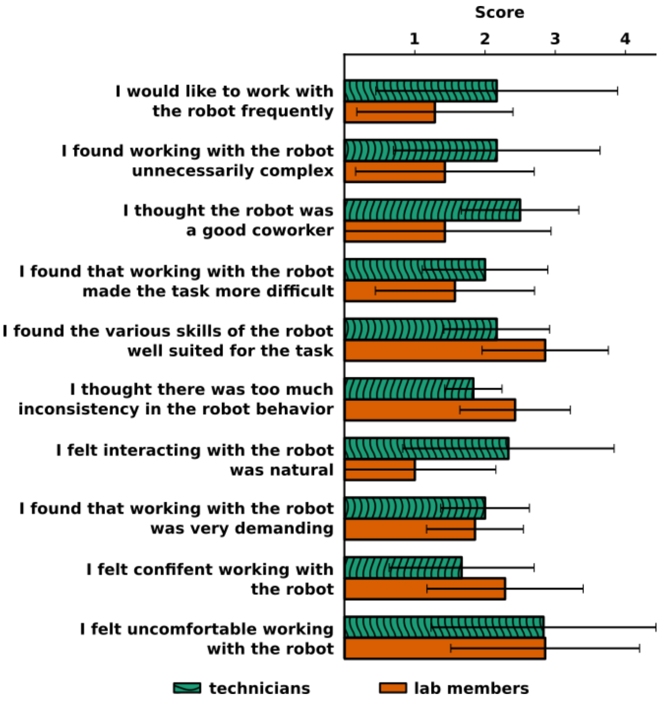


Fig. 5: Results of the modified SUS averaged over subjects of both groups. Scores are ranging from 0 to 4 with 4 the best value to obtain for each questions.

robots [29]. A plausible cause for this is that people have strong expectations regarding robots, which are not met when actually interacting with them. However, a direct comparison of the results is difficult due to the completely different settings (not the same robot nor the same task). In our setting, one plausible explanation lies in the constraints added to the interaction, e.g. having to use specific sentences to trigger an action, and the very scripted interaction. Previous research in HRI also show that the robots faults have an impact on subjects' subjective assessment [30], [31]. In our results, we found that to be true only for technicians and for the perceived safety measure. In [31] the authors specifically noticed a decrease in anthropomorphism when the robot is making a lot of errors. This effect does not seem to appear in our analysis but the small sample size makes it difficult to draw more conclusions. The perceived intelligence of the robot could probably be improved by adding more flexibility in the robot behavior, e.g. adapting the voice recognition software to include sentences with similar semantics.

Regarding correlations, it seems that for technicians, any traits that link the robot with a "human-like" partner are negatively correlated with usability. Nevertheless, those results need to be taken cautiously due to the very small sample size.

In terms of fixation patterns, there is a difference between the time spent doing visual servoing and looking at the hands of the robot compared to the human-human scenario. There could be multiple factors that explain those differences. First, time to perform the task was longer in the human-robot condition as the robot moves and acts slower than a human coworker. Visual servoing was mainly occurring when releasing the guard or placing it on the ground. When moving the guard with the partner there was almost no visual servoing happening. Therefore, it is not surprising that it represents a greater part of the task in the human-human condition as moving the guard take much longer with the robot. Interestingly, however, the time spent looking at the hands of the robot could suggest a need to ensure the robot had a good grip on the object, and, therefore, might be a metric for trust. Our hypothesis that differences between human-human and human-robot interactions would reduce with an increase in the number of repetitions is not validated by the data. A longer study with more subjects and more repetitions would be needed to validate any potential effects. Further experiments are required to fully link trust with this difference. Another interesting factor is time spent by the coworker looking at the head of the robot. In human interactions, head motion plays an important role as it conveys information to the partner [32]. On ARMAR-6, the head is of human form but not controlled in a human-like fashion. Currently, its sole purpose is to orient the head camera towards the object of interest or the human partner. The fact that subjects are looking at it during the interaction suggests that they might be looking for additional social cues. Therefore, controlling the head to convey such cues could be of interest.

Finally the free comments are in line with the results of the usability questionnaire. Adding more variability in the task, and in the robot behavior, should improve the usability of the system.

VI. CONCLUSION

The evaluation of ARMAR-6, a robotic assistant for industrial tasks, performed in an ecological environment, has shown that there is still room for improvement, especially in the robustness of the interaction and the usability of the whole system. The data we have gathered will be of prime importance to guide our further developments.

The number of failure events shows the necessity to improve the robustness of the interaction. Specifically, one event, the "re-grasping", shown in Fig. 3, introduced a few failure events. So did the voice triggering command as, due to the ambient noise in the Ocado factory, the subjects had to open and close the microphone before and after issuing a voice command. This was a safety measure that had to be applied, such that the robot did not unintentionally recognize background noises as a command, while already performing an action. However, this prevents us from using conversational interactions, e.g. the subject saying "Thank you" and the robot replying "You're welcome". Such conversation tools would enrich the interaction and makes it more natural. Therefore, solving both issues should drastically reduce the number of failure events and introduce a more natural interaction.

Analyzing the impact of the interaction on the gaze fixation patterns tends to demonstrate, first, that due to the anthropomorphic shape of the robot head, subjects are looking at it in search for social cues. Therefore, controlling the head for this purpose could be an interesting line of improvement. Second, the differences in terms of time spent looking at the robots hands might be a metric for trust. Further analysis and studies are necessary to validate these points.

Finally, it appears that there is a need for more standardized metrics to evaluate the interaction. The Godspeed questionnaire series is developed for this purpose but only serves to analyze the partners' subjective perception. The SUS technique, in its standard form, is not tailored for robotic evaluation as it is too generic. In this study we had to modified some of the questions which reduce possible comparison with other evaluations. A standardized evaluation metric for the usability of robotic system would be an essential tool for the community.

Acknowledgments

We would like to thanks all the partners of the Second-Hands consortium for providing the on-board algorithms used in this study and their feedback for this research.

References

[1] M. Pateraki and P. Trahanias, "Deployment of robotic guides in museum contexts," in *Mixed Reality and Gamification for Cultural Heritage*. Springer, 2017, pp. 449–472.

[2] A. Bilyea, N. Seth, S. Nesathurai, and H. Abdullah, "Robotic assistants in personal care: A scoping review," *Medical engineering & physics*, vol. 49, pp. 1–6, 2017.

[3] T. Shibata, "An overview of human interactive robots for psychological enrichment," *Proceedings of the IEEE*, vol. 92, no. 11, pp. 1749–1758, 2004.

[4] A. Sauppé and B. Mutlu, "The social impact of a robot co-worker in industrial settings," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 2015, pp. 3613–3622.

[5] A. Weiss and C. Bartneck, "Meta analysis of the usage of the godspeed questionnaire series," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 381–388.

[6] I. Maurtua, A. Ibarguren, J. Kildal, L. Susperregi, and B. Sierra, "Human–robot collaboration in industrial applications: Safety, interaction and trust," *International Journal of Advanced Robotic Systems*, vol. 14, no. 4, p. 1729881417716010, 2017.

[7] J. R. Campana and M. Quaresma, "The importance of specific usability guidelines for robot user interfaces," in *International Conference of Design, User Experience, and Usability*. Springer, 2017, pp. 471–483.

[8] S. Robla-Gómez, V. M. Becerra, J. R. Llata, E. Gonzalez-Sarabia, C. Torre-Ferrero, and J. Perez-Oria, "Working together: A review on safe human-robot collaboration in industrial environments," *IEEE Access*, vol. 5, pp. 26 754–26 773, 2017.

[9] C. L. Breazeal, *Designing Sociable Robots*. MIT press, 2004.

[10] B. Meerbeek, J. Hoonhout, P. Bingley, and J. M. Terken, "The influence of robot personality on perceived and preferred level of user control," *Interaction Studies*, vol. 9, no. 2, pp. 204–229, 2008.

[11] B. Reeves and C. Nass, "How people treat computers, television, and new media like real people and places," *CSLI Publications and Cambridge*, 1996.

[12] F. Kaplan, "Who is afraid of the humanoid? investigating cultural differences in the acceptance of robots," *International journal of humanoid robotics*, vol. 1, no. 03, pp. 465–480, 2004.

[13] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka, "Precision timing in human-robot interaction: coordination of head movement and utterance," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 131–140.

[14] R. P. van der Wel, G. Knoblich, and N. Sebanz, "Let the force be with us: dyads exploit haptic coupling for coordination." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 5, p. 1420, 2011.

[15] M. De Graaf, S. Ben Allouch, and J. Van Dijk, "Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 224–233.

[16] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[17] P. Boissy, S. Brière, H. Corriveau, A. Grant, M. Lauria, and F. Michaud, "Usability testing of a mobile robotic system for in-home telerehabilitation," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 1839–1842.

[18] D. Hebesberger, T. Koertner, C. Gisinger, and J. Pripfl, "A long-term autonomous robot at a care hospital: A mixed methods study on social acceptance and experiences of staff and older adults," *International Journal of Social Robotics*, vol. 9, no. 3, pp. 417–429, 2017.

[19] D. Y. Y. Sim and C. K. Loo, "Extensive assessment and evaluation methodologies on assistive social robots for modelling human–robot interaction–a review," *Information Sciences*, vol. 301, pp. 305–344, 2015.

[20] K. Werner, J. Oberzaucher, and F. Werner, "Evaluation of human robot interaction factors of a socially assistive robot together with older people," in *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*. IEEE, 2012, pp. 455–460.

[21] C. Gold, M. Körber, C. Hohenberger, D. Lechner, and K. Bengler, "Trust in automation–before and after the experience of take-over scenarios in a highly automated vehicle," *Procedia Manufacturing*, vol. 3, pp. 3025–3032, 2015.

[22] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[23] T. Asfour, L. Kaul, M. Wächter, S. Ottenhaus, P. Weiner, S. Rader, R. Grimm, Y. Zhou, M. Grotz, F. Paus, *et al.*, "Armar-6: A collaborative humanoid robot for industrial environments," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2018, pp. 447–454.

[24] M. Khoramshahi and A. Billard, "A dynamical system approach to task-adaptation in physical human–robot interaction," *Autonomous Robots*, vol. 43, no. 4, pp. 927–946, 2019.

[25] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[26] O. Friard and M. Gamba, "Boris: a free, versatile open-source event-logging software for video/audio coding and live observations," *Methods in Ecology and Evolution*, vol. 7, no. 11, pp. 1325–1330, 2016.

[27] J. Sauro, *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC Denver, CO, 2011.

[28] S. Keizer, P. Kastoris, M. E. Foster, A. Deshmukh, and O. Lemon, "Evaluating a social multi-user interaction model using a nao robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 318–322.

[29] K. S. Haring, Y. Matsumoto, and K. Watanabe, "How do people perceive and trust a lifelike robot," in *Proceedings of the world congress on engineering and computer science*, vol. 1, 2013.

[30] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.

[31] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 141–148.

[32] G. Charles, *Conversational organization: Interaction between speakers and hearers*. New York, Academic Press, 1981.