# Developing a Questionnaire to Evaluate Customers' Perception in the Smart City Robotic Challenge

Lun Wang* Luca Iocchi* Andrea Marrella* Daniele Nardi*

*Abstract*— **In this paper, we present an approach to develop a new type of questionnaire for evaluating customers' perceptions in the upcoming Smart CIty RObotic Challenge (SciRoc). The approach consists of two steps. First, it relies on interviewing experts on Human-Robot Interaction (HRI) to understand which robot's behaviours can potentially affect the users' perceptions during a HRI task. Then, it leverages a user survey to filter out those robot's behaviours that are not significantly relevant from the end user perspective. We concretely enacted our approach over a specific scenario developed in the context of SciRoc, which instructs a robot to take an elevator of a shopping mall asking support to the customers of the mall. The results of the survey have allowed us to derive a final list of 17 behaviours to be captured in the questionnaire, which has been finally developed relying on a 5-point Likert-scale.**

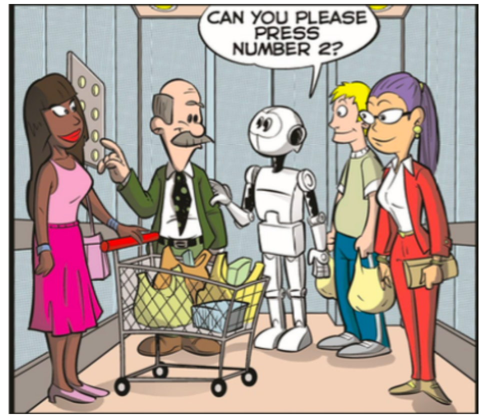



Fig. 1. Episode E4: "Take the Elevator"

## I. Introduction

The first Smart CIty RObotic Challenge (the SciRoc[1] challenge) will be held in Milton Keynes (UK) from 16 to 22 September 2019. SciRoc is a EU-H2020 funded project supporting the European Robotics League (ERL), and whose purpose is to bring all ERL tournaments in the context of smart cities. A key novelty of the SciRoc project is the introduction of robots in the ERL Smart Cities, in order to show how robots will integrate in the cities of the future as physical agents living in them. The challenge is focused on smart shopping and will be divided into a series of *episodes*, each consisting in a task to be performed through addressing specific research issues. Robots will be required to cooperate with the ICT infrastructure of Milton Keynes' (MK) shopping mall[2], whose "smartness" is given by a set of networked devices providing static and dynamic information from a number of heterogeneous data sources, e.g., location of shops, current availability of items, audio/visual inputs from CCTV cameras, crowd density sensor information, and many others. In addition, robots can interact with MK customers, accomplishing tasks of different nature in three different scenarios: assisting customers, providing professional services and supporting during emergency situations. Five episodes were finally designed and selected by the SciRoc consortium: (E3) Deliver coffee shop orders; (E4) Take the elevator; (E7) Serve goodies; (E10) Open the door; (E12) Fast delivery of emergency pills.

In this paper, we present an approach to develop a new type of questionnaire targeted to collect and evaluate customers' perceptions during social interactions between humans and robots in smart environments. To achieve this objective, we focused on the most social of the above episodes: (E4) Take the elevator . In E4, the robot must take an elevator of MK crowded with customers to reach a service located in another floor. The robot is able to enter/exit the elevator at the right floor in the presence of people nearby and/or inside. To perform the task, the robot can interact with the customers in spoken language. The robot is not supposed to push buttons, and it can ask the people around to do it (see Fig.1). According to the rules of E4, three to five users – randomly selected from the MK customers – will be involved to join the task performance: Two customers will have face-to-face interaction with the robot, while the others will be observers of the task. The duration of the task is 5 minutes.

To develop an effective questionnaire that is able to properly capture users' perceptions after their interaction with the robot in E4, it is necessary to understand which *robot's behaviours* may positively or negatively affect users' perceptions. To this end, a crucial step of our approach has consisted of identifying such behaviours through dedicated interviews and preliminary surveys performed with both experts and non-experts in the Human-Robotic Interaction (HRI) field, running two demos of the E4 episode. The details of the approach are presented in Section III, while the results of the interviews and surveys that provide the fundamentals of the questionnaire are discussed in sections IV and V. Based on such results, the concrete development of the questionnaire is shown in Section VI. Before describing the ingredients of our approach, we first provide in Section II some background information and related work necessary to

*Dipartimento di Ingegneria Informatica, Automatica e Gestionale "Antonio Ruberti", Sapienza Universitá di Rome, Rome, Italy Email: *lastname*@diag.uniroma1.it

[1]https://sciroc.eu/

[2]https://www.centremk.com/

motivate this research and understand the rest of the paper. Finally, in Section VII, we conclude the paper.

## II. Background and Related Work

In the HRI field, *query techniques* are used to collect details of the user's perception about the interaction with a robot. The advantage of such methods is that they get the user's viewpoint directly and may reveal issues that have not been considered by the designer [1]. Two main types of query techniques are used in HRI: *interviews* and *questionnaires*.

On the one hand, interviewing users about their experience with an interactive robot provides a direct and structured way of gathering information. Interviews have the advantage that the level of questioning can be varied to suit the context. Interviews are particularly effective for collecting information about user preferences, impressions and attitudes. On the negative side, they are time consuming; consequently, they usually reach a limited number of users.

On the other hand, an alternative method of querying the user is to administer a questionnaire [1]. In questionnaires, since questions are fixed in advance, the level of "flexibility" is lower than in interviews. Nonetheless, they can be used to reach a wider participant group as they take less time to administer, and the answers can be analyzed more rigorously. Since the evaluator is not directly involved in the completion of the questionnaire, it is crucial that it is well designed. First, the evaluator must establish the purpose of the questionnaire. Then, it is useful to decide how the questionnaire responses are to be analyzed (e.g., measurable feedbacks, users' impression, etc.). The latter depends on the styles of question that are included in the questionnaire: *(i) Open-ended questions* ask the users to provide their own opinions; they are useful for gathering subjective information but are complex to analyze in any rigorous way; *(ii) Scalar questions* ask the user to judge a specific statement on a numeric scale (e.g., Likert scale ranging from 1 to 5), which usually corresponds to a measure of agreement or disagreement with the statement; *(iii) Multi-choice questions* offer the user with the choice to select one or many explicit responses.

To date, within the context of existing query techniques employed in HRI, there is no clear method that covers the breadth and depth of the socially driven interaction experience with robots in smart environments. Some query techniques exist, especially in form of questionnaires, but they cover *generic aspects* of the interaction [2]. For example, Godspeed questionnaire uses a Likert scale for measuring the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of robots [3]. The Robotic Social Attributes Scale (RoSAS) has been developed and validated to measure social perception of robots [4]. The Negative Attitude toward Robots Scale (NARS) measures humans' attitudes toward communication robots in daily-life [5]. Finally, the Subjective Assessment of Speech System Interfaces (SASSI) provides a questionnaire to measure the subjective evaluation of speech recognition interfaces [6].

As a matter of fact, there is a lack of frameworks that synthesize the various existing query methods together to target not only the social nature of HRI, but also the wider contexts within which the interaction happens. To tackle this challenge, in this paper we realize a simple and repeatable approach to the development of a questionnaire that is specifically tailored for the context where the interaction happens. In particular, we leverage the scenario presented in E4 to develop a *new type of questionnaire* that refines and integrates the existing methods focusing on three perspectives, which describe interaction experiences with robots in smart environments[3]: *Social behavior of robot, Proxemics between human and robot, Collaboration behavior of robot*.

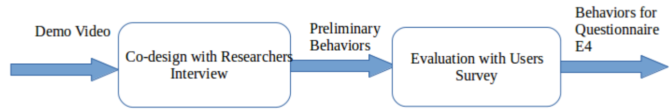



Fig. 2. Overview of the Approach

## III. Approach

The proposed approach to the questionnaire development is based on two steps, as shown in Fig. 2. First, we relied on dedicated interviews with researchers in robotics to understand which robot's behaviours can potentially affect the users' perceptions during a task of HRI in a smart environment. Then, leveraging a user survey performed with non-experts in HRI, we validated which of the previously selected behaviours are concretely relevant from the user perspective, filtering out those ones considered as secondary. The results of this second step have allowed us to derive the final list of behaviours to capture in the questionnaire.

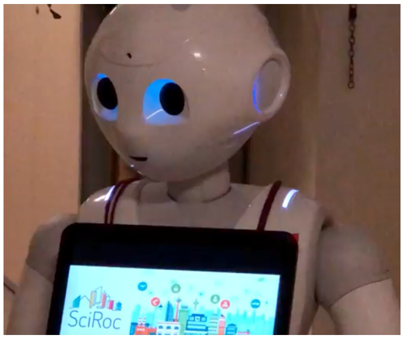
### A. Role of users in E4

To understand our approach, we first need to explain the role of users in the episode E4 under investigation. According to the rules defined for E4, the robot will encounter two persons while moving towards the elevator: Person A stands in a location and s/he is not interested in interacting with the robot; Person B is actively moving towards the robot willing to interact with it. Once arrived in front of the elevator, the robot will encounter between one to three persons (Person C, D and E). All such people will take the elevator with the robot, but Person C is the only one that can interact with the robot inside the elevator to support it in pushing the button for the right floor. All the other persons (A, D and E) will passively observe the performance of the task until its completion, i.e., they will not interact with the robot. Thus, we consider them as *observers* of the task.

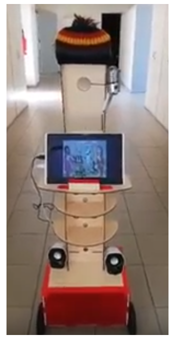
### B. Implementation of two demos for E4

To capture the potential interesting behaviours of a the robot during E4, we implemented, executed and recorded

[3]The three perspectives were selected by the SciRoc Consortium to evaluate the interaction experiences with robots in the various episodes.



(a) Pepper



(b) MARRtino

Fig. 3. Social robots selected to perform the two demos

(through a video camera) two complete demos of the episode employing two different types of robot.

For the first demo, we chose Pepper[4] as social robot. Pepper (cf Fig.3(a)) is a human-like service robot produced by SoftBank, which can interact with users through spoken language or, alternatively, with a tablet attached to the robot. The tablet displays images and allows for tactile interaction.

For the second demo, we chose MARRtino[5] as social robot. MARRtino (cf. Fig.3(b)) is a ROS-based low-cost differential drive robot platform that can be assembled in many shapes. We assembled MARRtino with a tablet, a microphone and a loudspeaker. The tablet displays images, detects faces, and communicates with users.

We performed both demos using Wizard-of-Oz method [7]. The video recordings of the demos are available online, at: *https://bit.ly/2I8zBxU* and *https://bit.ly/2VhGvbT*. It is worth to notice that our decision to employ two robots having different shapes (Pepper with human-like features, MARRtino with a machine-like shape) has been targeted to avoid that the shape of the robot could influence the spectrum of behaviours perceived by the users during a HRI task.

## IV. Results of the Interviews

Recording demos of E4 was useful to show the concrete working of the episode to several users. In this direction, we realized dedicated interviews with five researchers of the robotic laboratory at Sapienza University of Rome, showing them the video demos of E4. The interviews were targeted to investigate which of the behaviours covered by the existing query methods in HRI are able to potentially affect the users' perceptions during E4. To this end, we grouped such behaviours according to the three perspectives discussed at the end of Section II, and we analyzed them one by one.

**1) Social behavior of robot.** For this perspective, we relied on RoSAS [4], which was developed exclusively to measure the social behavior of a robot in a HRI task. RoSAS considers the following behaviours:

- **Warmth of robot**: *Happy, Feeling, Social, Organic, Compassionate, Emotional*.
- **Competence of robot**: *Capable, Responsive, Interactive, Reliable, Competent, Knowledgeable*.
- **Discomfort of robot**: *Scary, Strange, Awkward, Dangerous, Awful, Aggressive*.

The researchers involved in the interview evaluated all the behaviours covered by RoSAS as potentially relevant (except *Organic*, which was filtered out) for measuring users' perception in E4.

**2) Proxemics between human and robot.** Proxemics is the study of how a robot uses the available space considering the physical and psychological distancing from humans [8]. People might perceive robots that do not show appropriate distance as threatening to their social environments and work practices [9]. On the other hand, carefully designed proxemic behaviors in robots might foster closer relationships with humans, enabling widespread acceptance of robots. While this issue is well investigated in the literature (e.g., see [10], [11]), in E4 we decided to analyze proxemics through customers' perceptions. In this direction, the researchers involved in the interview considered the following behaviours, covered by [3] and [12], as potentially relevant in E4, except the *competence of robot*, which was already discussed in the previous perspective:

- **Engagement**: *Robot looks at an object, robot looks at user's eyes, Robot looks at user's face, users look at robot's face, users look at robot's eyes, users pay attention to the conversation with robot, users understand well the meaning of the conversation*.
- **Anthropomorphism of robot**: *Natural, Human-like, Conscious, Lifelike, Moving elegantly*.
- **Likeability of robot**: *Like, Friendly, Kind, Pleasant, Nice*. Moreover, researchers suggested the insertion of an additional behaviour: *Politeness* of the robot.
- **Technology adoption by customers**: *Adaptability, ease of use, usefulness, trust*.

**3) Collaboration with robot.** On the basis of previous studies on collaborative attitudes of humans towards robots (see [11], [13], [14]), the following behaviours were considered as potentially relevant in E4 by the researchers involved in the interview:

- **Users' willingness to help**: *Attractiveness, Enjoyment, Endearment, Symbiotic relationship, Reciprocity relationship, Collaborativeness*.

Conversely, researchers decided to exclude the behaviours related to the avoidance of users to help, e.g., *Unconcerned, Authoritative, Handful, Hateful*.

## V. Results of the Survey

After the identification of preliminary robot's behaviours through the interviews, we performed a survey with 100 users having no previous experience of HRI. Such users were invited to watch the video recordings of the two demos: 51 of them watched the first demo, the other 49 watched the second demo. Users were selected randomly among the undergraduate students at Sapienza University of Rome. Once a specific user completed watching the video recording

[4]https://www.softbankrobotics.com/emea/en/robots/pepper
[5]https://www.marrtino.org

TABLE I

SOCIAL BEHAVIOR OF ROBOT (DEMO 1)

| | k | 1st step | 2rd step(P-value) | 3rd step |
|---|---|---|---|---|
| ***Warmth of robot*** | | | | |
| Happy | 23 | - | 0,576 ✓ | ✓ |
| Feeling | 2 | ✗ | - | - |
| Social | 44 | ✓ | - | ✓ |
| Compassionate | 1 | ✗ | - | - |
| Emotional | 1 | ✗ | - | - |
| ***Competence of robot*** | | | | |
| Capable | 11 | - | - | 0,000 ✗ |
| Responsive | 26 | ✓ | - | ✓ |
| Interactive | 39 | ✓ | - | ✓ |
| Reliable | 2 | ✗ | - | - |
| Competent | 3 | ✗ | - | - |
| Knowledgeable | 17 | - | 0,024 ✗ | - |
| ***Discomfort of robot*** | | | | |
| Scary | 2 | ✗ | - | - |
| Strange | 26 | ✓ | - | ✓ |
| Awkward | 13 | - | 0,001 ✗ | - |
| Dangerous | 1 | ✗ | - | - |
| Awful | 1 | ✗ | - | - |
| Aggressive | 0 | ✗ | - | - |

TABLE II

SOCIAL BEHAVIOR OF ROBOT (DEMO 2)

| | k | 1st step | 2rd step(P-value) | 3rd step |
|---|---|---|---|---|
| ***Warmth of robot*** | | | | |
| Happy | 12 | - | 0,000 ✗ | - |
| Feeling | 5 | ✗ | - | - |
| Social | 42 | ✓ | - | ✓ |
| Compassionate | 1 | ✗ | - | - |
| Emotional | 1 | ✗ | - | - |
| ***Competence of robot*** | | | | |
| Capable | 18 | - | 0,085 ✓ | ✓ |
| Responsive | 30 | ✓ | - | ✓ |
| Interactive | 31 | ✓ | - | ✓ |
| Reliable | 5 | ✗ | - | - |
| Competent | 5 | ✗ | - | - |
| Knowledgeable | 12 | - | 0,000 ✗ | - |
| ***Discomfort of robot*** | | | | |
| Scary | 0 | ✗ | - | - |
| Strange | 33 | ✓ | - | ✓ |
| Awkward | 16 | - | 0,021 ✗ | - |
| Dangerous | 0 | ✗ | - | - |
| Awful | 0 | ✗ | - | - |
| Aggressive | 0 | ✗ | - | - |

of a demo, s/he was asked (through the survey) to indicate which robot's behaviours s/he perceived as relevant during the recorded HRI task. In a nutshell, for any of the behaviours identified in Section IV, the user could decide to select (or neglect) it, e.g., a user could consider the interaction as *Social* but not *Knowledgeable*. We finally collected 100 answers to the survey. The raw data of the survey is available online at: `http://tiny.cc/6v2b6y`.

### A. Data Analysis

The gender distribution was as follows: 74% of male users, 24% of female users and 2% of users that did not declare their gender. The average age of users was 21,05 years old.

The analysis of the data collected with the survey has been enacted following a three-steps approach.

- First, for each demo, we assessed how many times a behaviour was selected by the users. Specifically, if $N$ is the amount of performed surveys for a specific demo and $K$ is the number of times that a behaviour $B$ is selected, we decided to consider $B$ as relevant for the final questionnaire if $K/N > 0.5$, i.e., if more than half of the users selected $B$ in the survey. Conversely, we immediately discarded all behaviours for which $K/N < 0.2$.
  For example, if we consider the first demo, the *Social* behaviour of the robot was selected $K = 44$ times out of $N = 51$ surveys (see Table I). This means that more than 86% (i.e., 44/51 = 0.86) of users perceived the robot as *Social*. On the other hand, only 1 user perceived the robot as *Dangerous*, meaning that only 1.9% (i.e., 1/51 = 0.019) of users selected this behaviour as relevant. Consequently, it was filtered out by the questionnaire.
- Secondly, for all those behaviours for which $0.2 \leq K/N \leq 0.5$, we performed a statistical test to investigate if it could be considered for the questionnaire or not. Specifically, we leveraged the *binomial test* [15], a statistical procedure that compares the observed frequencies of the two categories of a dichotomous variable[6] to the frequencies that are expected under a binomial distribution with a specified probability parameter. The *null hypothesis* for this kind of test, which is also the one we wanted to confirm, is that the observed results do not differ significantly from what is expected. To be properly run, a binomial test requires three ingredients: $N$, $K$, and a probability parameter, that in our case can be set to 0.5.
  For example, if we consider the first demo, 23 users perceived the robot as *Happy*, i.e., the observed proportion was $K/N = 23/51 = 0.45$. Our null hypothesis states that this proportion should be 0.5 for the entire population of $N = 51$ surveys. The outcome of the binomial test is the *p value*, which in this case is 0.576, i.e., there is 57.6% of chance of selecting 23 times the behavior *Happy* out of N = 51 surveys performed with respective users. If this chance had been smaller than 5% ($p < 0.05$), the null hypothesis would have been rejected. In this example, the null hypothesis is confirmed, meaning that *Happy* can be considered as a relevant behaviour for the questionnaire. Conversely, on the other hand, if we consider the second demo (see Table II), we found that 16 users perceived the robot as *Awkward* out of $N = 49$ surveys, i.e., the observed proportion was $K/N = 16/49 = 0.32$. In this case, the computed $p$ value was 0.021, meaning that there is 2.1% of possibilities that *Awkward* is selected 16 times in 49 experiments. Since $p < 0.05$, the null hypothesis is rejected. Consequently, this behaviour was filtered out by the questionnaire.
  It is worth to notice that a behaviour that successfully passed the first or second step of the approach in one demo but not in the other was considered as sufficiently

[6] A dichotomous variable is a variable that can take only two possible values: yes or no, true or false, 0 or 1, and so on.

relevant for our purposes, meaning that it was inserted in the final questionnaire. This was the case of *Robot looks at user's face*, which did not pass the first and second step in the second demo (see Table IV), but passed the first step in the first demo (see Table III).

- Third, we used the *Pearson correlation* to investigate if it was possible to merge two different behaviours in a single one. In a nutshell, a Pearson correlation is a measure of the strength of the linear relationship between two dichotomous variables. The outcome is a number $r$ that varies between -1 and 1, reflecting the extent to which two variables are linearly related. A value of $r$ equal to -1/1 indicates a perfect negative/positive linear relationship between the two investigated variables. On the other hand, if $r$ is equal to 0, there is no linear relationship between two variables. In our case, for all those behaviours who "survived" from the first two steps of the approach, we decided to apply the Pearson correlation and check if they were strongly (positively or negatively) correlated, i.e., with $r > 0.5$ or $r < -0.5$, or not. Just in one case we found a strong positive correlation. Analyzing the results of the surveys related to the first demo, under the *Proxemics between human and robot* perspective (see Table III) we found that behaviours *Users pay attention to conversation with robot* and *Users understand well the meaning of conversation* were strongly related in a positive way ($r = 0.704$). However, since this same correlation was not verified also in the second demo ($r > 0$ but $r < 0.5$, see Table IV), we finally could not merge the two behaviours in a single one.

The results of the surveys analyzed through our three-steps approach are shown in tables I, II, III, IV, V and VI. Results are categorized per demo and according to the considered perspective.

TABLE III
PROXEMICS BETWEEN HUMAN AND ROBOT (DEMO 1)

| | k | 1st stp | 2rd stp(p) | 3rd stp |
|---|---|---|---|---|
| ***Engagement*** | | | | |
| Robot looks at an object | 11 | - | 0,000 ✗ | - |
| Robot looks at user's eyes | 9 | ✗ | - | - |
| Robot looks at user's face | 33 | ✓ | - | ✓ |
| Users look at robot's eyes | 11 | - | 0,000 ✗ | - |
| Users look at robot's face | 22 | - | 0,401 ✓ | ✓ |
| Users pay attent. to conversa. | 18 | - | 0,049 ✓ | ✓ |
| Users understand well conversa. | 19 | - | 0,092 ✓ | ✗ |
| ***Anthropomorphism of robot*** | | | | |
| Natural | 11 | - | 0,000 ✗ | - |
| Human-like | 11 | - | 0,000 ✗ | - |
| Conscious | 12 | - | 0,000 ✗ | - |
| Lifelike | 11 | - | 0,000 ✗ | - |
| Moving elegantly | 10 | ✗ | - | - |
| ***Likeability of robot*** | | | | |
| Like | 10 | ✗ | - | - |
| Friendly | 27 | ✓ | - | ✓ |
| Kind | 16 | - | 0,011 ✗ | - |
| Pleasant | 6 | ✗ | - | - |
| Nice | 16 | - | 0,011 ✗ | - |
| Politeness | 32 | ✓ | - | ✓ |
| ***Technology adoption*** | | | | |
| Adaptability | 8 | ✗ | - | - |
| Ease of Use | 23 | - | 0,576 ✓ | ✓ |
| Useful | 14 | - | 0,002 ✗ | - |
| Trust | 11 | - | 0,000 ✗ | - |

TABLE IV
PROXEMICS BETWEEN HUMAN AND ROBOT (DEMO 2)

| | k | 1st stp | 2rd stp(p) | 3rd stp |
|---|---|---|---|---|
| ***Engagement*** | | | | |
| Robot looks at an object | 5 | ✗ | - | - |
| Robot looks at user's eyes | 1 | ✗ | - | |
| Robot looks at user's face | 12 | - | 0,000 ✗ | - |
| Users look at robot's eyes | 1 | ✗ | - | - |
| Users look at robot's face | 17 | - | 0,044 ✓ | ✓ |
| Users pay attent. to conversa. | 22 | - | 0,568 ✓ | ✓ |
| Users understand well conversa. | 23 | - | 0,775 ✓ | ✓ |
| ***Anthropomorphism of robot*** | | | | |
| Natural | 4 | ✗ | - | - |
| Human-like | 2 | ✗ | - | - |
| Conscious | 32 | ✓ | - | ✓ |
| Lifelike | 5 | ✗ | - | - |
| Moving elegantly | 3 | ✗ | - | - |
| ***Likeability of robot*** | | | | |
| Like | 2 | ✗ | - | - |
| Friendly | 23 | - | 0,775 ✓ | ✓ |
| Kind | 15 | - | 0,009 ✗ | - |
| Pleasant | 9 | ✗ | - | - |
| Nice | 8 | ✗ | - | - |
| Politeness | 24 | - | 1,000 ✓ | ✓ |
| ***Technology adoption*** | | | | |
| Adaptability | 18 | - | 0,085 ✓ | ✓ |
| Ease of Use | 25 | ✓ | - | ✓ |
| Useful | 16 | - | 0,021 ✗ | - |
| Trust | 6 | ✗ | - | - |

## VI. IMPLEMENTING THE QUESTIONNAIRE FOR E4

The results of the survey allowed us to identify the 17 behaviours to include in the questionnaire, which will be delivered during E4 in the context of the SciRoc challenge. The questions were organized as follows, according to the perspective they belong to:

**Social Behavior of robot**

- Have you perceived happiness of the robot?
- Have you perceived sociability of the robot?
- Have you perceived capability of the robot?
- Have you perceived responsiveness of the robot?
- Have you perceived interactiveness of the robot?
- Have you perceived strangeness of the robot?

**Proxemics between human and robot**

- Did robot look at users' face during the conversation between user and the robot?
- Did user look at the robot's face during the conversation between user and the robot?
- Have users paid attention to the conversation with the robot?

- Have users understood well the meaning of conversation?
- Have you perceived consciousness of the robot?
- Have you perceived friendliness of the robot?
- Have you perceived politeness of the robot?
- Have you perceived adaptability of the robot?
- Have you perceived ease of use with the robot?

**Collaboration with robot**

- Have you perceived enjoyment of the robot?
- Have you perceived collaborativeness of the robot?

Any question can be answered selecting one item from a 5-point Likert scale [16]. The scores assigned the scale can vary on the basis of the specific behaviour under investigation. Specifically, if the robot's behaviour involved in a specific question reflects a positive attitude, answers are of kind: *Absolutely No=1, No=2, Neutral=3, Yes=4, Absolutely Yes=5*. Conversely, in case of a negative attitude, e.g., for a negative behavior like *Strange*, the scores assigned to the answers are turned over, i.e., *Absolutely No=5, No=4, Neutral=3, Yes=2, Absolutely Yes=1*.

Moreover, for any of the 17 questions, we added a "*not available*" additional answer. Users can select it when a specific HRI task is not completely performed from the start to the end, or when a specific behavior of a robot is not perceived during the concrete interaction with it.

## VII. Conclusion

In the end, the complete questionnaire that we developed to evaluate the performance of robots in E4 at the first SciRoc challenge is available online at: *http://tiny.cc/vqei6y*. This questionnaire will be provided to the teams and discussed in the coming rule discussion period. However, we believe that it stands as a novel contribution to the evaluation of performance of robots in social and smart contexts. Moreover, we consider our approach to design the questionnaire as a first step towards systematic evaluation of robots performances in competition settings that specifically address HRI issues.

The immediate future work is to challenge the robustness of our approach and the effectiveness of the proposed questionnaire not only by administering it to a larger and more diverse set of users during the enactment of the episode E4 in the range of the SciRoc challenge, i.e., in a real smart city context with robots, but also exploiting it for collecting data from the other SciRoc episodes, to validate its ability to generalize towards different yet similar contexts.

TABLE V
Collaboration with robot (Demo 1)

| | k | 1st step | 2rd step (p-value) |
|---|---|---|---|
| ***Users' willingness to help*** | | | |
| Attractiveness | 9 | ✗ | - |
| Enjoyment | 7 | ✗ | - |
| Endearment | 16 | - | 0,011 ✗ |
| Symbiotic Relationship | 7 | ✗ | - |
| Reciprocity Relationship | 10 | ✗ | - |
| Collaborativeness | 27 | ✓ | - |

TABLE VI
Collaboration with robot (Demo 2)

| | k | 1st step | 2rd step(p) | 3rd step |
|---|---|---|---|---|
| ***Users' willingness to help*** | | | | |
| Attractiveness | 9 | ✗ | - | - |
| Enjoyment | 18 | - | 0,085 ✓ | ✓ |
| Endearment | 2 | ✗ | - | - |
| Symbiotic Relationship | 6 | ✗ | - | - |
| Reciprocity Relationship | 9 | ✗ | - | - |
| Collaborativeness | 27 | ✓ | - | ✓ |

## References


[1] A. Dix, J. Finlay, G. Abowd, and R. Beale, "Human-Computer Interaction," *Pearson*, 2004.
[2] J. E. Young, J. Sung, A. Voida, E. Sharlin, T. Igarashi, H. I. Christensen, and R. E. Grinter, "Evaluating human-robot interaction," *Int. Journal of Social Robotics*, vol. 3, no. 1, 2011.
[3] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. Journal of social robotics*, vol. 1, no. 1, 2009.
[4] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas): Development and validation," in *2017 ACM/IEEE Int. Conf. on Human-Robot Interaction*. ACM, 2017.
[5] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots," in *13th IEEE Int. Workshop on Robot and Human Interactive Communication*. IEEE, 2004.
[6] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (sassi)," *Natural Language Engineering*, vol. 6, no. 3-4, 2000.
[7] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, 2012.
[8] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *6th Int. Conf. on Human-Robot Interaction*. ACM, 2011.
[9] P. Papadakis, P. Rives, and A. Spalanzani, "Adaptive spacing in human-robot interactions," in *2014 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2014.
[10] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *2009 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2009.
[11] F. Riccio, A. Vanzo, V. Mirabella, T. Catarci, and D. Nardi, "Enabling symbiotic autonomy in short-term interactions: A user study," in *Int. Conf. on Social Robotics*. Springer, 2016.
[12] R. Mead and M. J. Matarić, "Proxemics and performance: Subjective human evaluations of autonomous sociable robot distance and social signal understanding," in *2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2015.
[13] M. Bajones, A. Weiss, and M. Vincze, "Help, anyone? a user study for modeling robotic behavior to mitigate malfunctions with the help of the user," *arXiv preprint arXiv:1606.02547*, 2016.
[14] A. Weiss, J. Igelsböck, M. Tscheligi, A. Bauer, K. Kühnlenz, D. Wollherr, and M. Buss, "Robots asking for directions: the willingness of passers-by to support robots," in *5th ACM/IEEE Int. Conf. on Human-Robot Interaction*. IEEE Press, 2010.
[15] K. A. Brownlee, *Statistical theory and methodology in science and engineering*. Wiley New York, 1965, vol. 150.
[16] J. Dawes, "Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales," *Int. Journal of Market Research*, vol. 50, no. 1, 2008.