A Methodology to Design and Evaluate HRI Teaming Tasks in Robotic Competitions

ANDREA MARRELLA, LUN WANG, LUCA IOCCHI, and DANIELE NARDI,
Sapienza University of Rome, Italy

As social robots become more prominent in our lives, their interaction with humans takes an increasing role, and new collaborative scenarios emerge. This development brings the need to realize robust test methods enabling the design and evaluation of Human-Robot Interaction (HRI) teaming tasks to prove functionality and promote adoption. In this article, we present a general-purpose and repeatable methodology for conducting studies in collaborative HRI in the range of robotic competitions. The methodology includes a step-by-step approach to design HRI teaming tasks tailored to be enacted in a robotic competition and to evaluate the performance of social robots to execute the designed tasks, exploring the relationship between robots' performance and user perceptions based on the feedback of the users participating to such tasks. We assess the feasibility of the methodology to design and evaluate an HRI teaming task in the context of "Smart CIties RObotics Challenges" (SciRoc) competition, which targets at investigating the impact of social of robots in smart cities.

CCS Concepts: • **Human-centered computing → Interaction design process and methods**; **Interaction design theory, concepts and paradigms**; • **Computer systems organization → Robotics;**

Additional Key Words and Phrases: Methodology, design of HRI teaming task, HRI evaluation, robotic competition, SciRoc

ACM Reference format:
Andrea Marrella, Lun Wang, Luca Iocchi, and Daniele Nardi. 2022. A Methodology to Design and Evaluate HRI Teaming Tasks in Robotic Competitions. *Trans. Hum.-Robot Interact.* 11, 3, Article 34 (September 2022), 23 pages.
https://doi.org/10.1145/3528415

1 INTRODUCTION

Human-Robot Interaction (**HRI**) is a rapidly advancing area of research in **Artificial Intelligence** (**AI**), dedicated to understanding, designing, and evaluating robotic systems for use by or with humans [18]. In the last years, advances in AI have led to robots endowed with sophisticated

Andrea Marrella and Lun Wang contributed equally to the article.

The work of Andrea Marrella has been partially supported by European Research Council under the European Unions Horizon 2020 Program through the DATACLOUD, DESTINI and FIRST projects. The work of Lun Wang, Luca Iocchi and Daniele Nardi has been supported by European Research Council under the European Unions Horizon 2020 Program through the SciRoc and AI4EU projects.

Authors' address: A. Marrella, L. Wang, L. Iocchi, and D. Nardi, Sapienza University of Rome, via Ariosto 25, 00185, Rome, Italy; emails: {marrella, wang, iocchi, nardi}@diag.uniroma1.it.

2573-9522/2022/09-ART34 $15.00
https://doi.org/10.1145/3528415

social abilities, which promise to enhance our daily lives [20]. As social robots become increasingly present in human society, for understanding how best to introduce them to complex and collaborative application domains, such as in schools, hospitals, and at home, there is a growing need for novel and robust methods of evaluation that offer metrics and tools to assess how humans respond to and collaborate with robots, how they feel about their interactions with robots, and how they interpret the actions of robots [5].

Despite the HRI literature provides many evaluation methods for performing HRI studies in collaborative scenarios (sometimes similar to the ones available in the Human-Computer Interaction field [21, 22, 31]), such as interviews and questionnaires [14], behavioral measures [33], psychophysiology measures [10] and task performance metrics [50], the reliable assessment of the HRI performance and social abilities of robots requires performing experiments under *replicable* conditions, in order to establish traceable mechanisms for vendors and consumers of HRI technologies to assure functionality [32]. In this direction, *robotic competitions* provide good opportunities for conducting empirical research studies in the field of collaborative HRI, because among their underpinning principles, replicability is considered crucial for a rigorous evaluation and comparison of results [2].

The idea of leveraging robotic competitions as experimental procedures were already proved as a valuable tool to perform HRI evaluation of social robots [2, 24, 34, 38, 59, 62, 63]. Nonetheless, a clear methodology that delineates the steps to design and evaluate HRI teaming tasks suitable to be executed in the range of a scientific competition is currently missing in the research literature. This lack of methodological guidance has led previous research works to often adopt the **Wizard-of-Oz** (**WoZ**) method to perform HRI analysis in robotic competitions, i.e., with an expert user acting "behind-the-scene" that controls the movements and interactions of the robot. However, according to [43], a WoZ controlled robot serves more as a proxy for a human and less as an independent entity, thus transforming HRI into a human-human interaction via a robot, which may negatively bias the assessment of collaborative HRI.

In this article, we address this challenge by presenting a general-purpose and repeatable methodology for conducting studies in collaborative HRI in the range of robotic competitions. Our methodology includes a step-by-step approach to:

- *design* HRI teaming tasks tailored to be executed in a robotic competition by fully autonomous robots, i.e., robots that are configured to act autonomously without the need of any external guidance. If compared with previous works in this area [2, 24, 34, 38, 59, 62, 63], here a strong focus is provided on the detailed specification of the collaborative scenario where the HRI task is intended to take place;
- *evaluate* the performance of social robots to execute the designed tasks for exploring the relationship between robots' performance and user perceptions based on the feedback of the users participating in such tasks. The target is to understand more about users' perception of an HRI teaming task, which is considered as a key driver to enable a social robot to adapt its behavior with respect to the users' characteristics and preferences [44].

To assess the feasibility of our methodology, we describe the specific tools and techniques employed for operationalizing it to design and evaluate an HRI teaming task in the context of the first "**Smart CIties RObotics Challenges**" (**SciRoc**) competition, whose target was to demonstrate how social robots can be useful to the customers of a shopping mall.

Specifically, we leveraged our methodology to design an HRI teaming task in which a robot is instructed to take an elevator of the shopping mall asking for customers support to achieve its objective. We selected a representative sample of real customers as active participants in the

task, having no background in robotics. Five teams, providing autonomous robots having slightly different appearances and interfaces, were involved in the task execution.

Concerning the evaluation, we first assessed the robots' performance according to what the teams/robots achieved (and not achieved) during the task execution, assigning them a score leveraging on a fair judging system. Then, we asked the users participating in the task to fill a dedicated questionnaire [56] for investigating their perception of the robots' behavior. In particular, we performed a confirmatory research study to validate three research hypotheses related to the impact of the *robots' interaction modalities*, *users' gender*, and *users' role* in the robots' behaviors as perceived by the users participating in the task. Then, we performed an exploratory research study to analyze the relationship between the measured robots' performance and their perceived behavior by users to search for interesting insights.

The results indicate that: (i) even in the case of robots having an almost identical appearance, slightly changing the interaction modalities can affect how the users' perceive the robots' behavior; (ii) only some social behaviors of robots are concretely influenced by the users' gender; (iii) the users' role has an impact on the users' perception of robots' behavior only when the interaction with the robots is conducted in spoken language; (iv) robots' behaviors perceived by users can be (sometimes) predictors of robots' performance in executing an HRI teaming task, in particular for the behaviors related to *Perceived Interactiveness* and *Perceived Collaborativeness*.

The rest of the article is organized as follows. Section 2 discusses previous literature works that propose generalized HRI frameworks leveraging robotic competitions, and explains the novel research aspects addressed in this article. Section 3 presents our general purpose and repeatable methodology for conducting HRI studies in the range of robotic competitions. Section 4 shows an instantiation of the methodology over a real robotic competition SciRoc, to show its feasibility to design and evaluate a concrete HRI task. Section 5 reports the results obtained from the evaluation of the HRI task designed through our methodology. A discussion on handling external influencing factors neglected in the evaluation is also included to account for the results achieved. Finally, Section 6 draws future work and conclusions.

2 RELATED WORK

The idea of leveraging robotic competitions as a tool to evaluate the outcome of HRI tasks was originally proposed in 2004 by Yanco et al. [62]. Specifically, the authors studied the videotapes of four different robot systems performing urban search and rescue tasks in a controlled environment in the range of the 2002 AAAI Robot Rescue Competition. Starting from the analysis of the videotapes of the robots in action, their user interfaces and the behavior of human operators involved in the tasks, a set of preliminary guidelines for the design of interfaces for HRI was defined.

In [59], Weiss et al. conducted a study in the context of the ICRA 2008 HRI Challenge. Five team's took part in the competition. Each team showed iteratively to a physical audience a live demo of a robot interacting with the teams' members during a specific HRI task. A panel of seven experts, as well as the attendees who had the opportunity to see the demos, assessed the demonstrations according to the robot's social and learning skills. The evaluation was based on a standardized questionnaire. Not surprisingly, the evaluation revealed the subjectivity of people's perception of HRI, emphasizing the role that the country of origin and level of expertise had on the overall assessment.

In [63], the authors exploited the **DARPA Robotics Challenge** (**DRC**) Trials enacted in 2013 to investigate the performance of HRI tasks in disaster response scenarios. Each of the eight teams participating in the study was involved in HRI tasks with a humanoid robot (acting on the field) and many operators located in a control room. Performance metrics, such as incidents and utterances, were analyzed categorizing them on the basis of the number of operators involved in the tasks

and the interaction/control methods employed. The study was useful to confirm that, in the case of HRI tasks involving unmanned ground vehicles, fewer operators and more automation lead to better performance. Similarly, Norton et al. [38] conducted an HRI study at the DRC 2015 Finals competition, which involved 20 teams consisting of human operators located in a control room and humanoid robots executing a variety of HRI tasks on the field. The results of the study were extrapolated to delineate recommendations to design of HRI tasks with remote humanoids.

In [34], Mizuchi and Inamura investigated the extent to which the results of traditional subjective evaluations (e.g., through questionnaires) performed to assess HRI tasks can be approximated using objective factors. The authors analyzed HRI history data coming from a robot competition task in which the robot was required to generate comprehensible natural language expressions to guide inexpert users in a virtual reality environment. The outcomes of the analysis revealed that the subjective evaluation of HRI can be reasonably approximated based on objective factors.

Differently from the works [34, 38, 63], which focused on evaluating HRI tasks performed in a virtual/remote setting (e.g., relying on the WoZ method), in [24] Iocchi et al. presented the RoboCup@Home competition, where domestic service robots execute several tasks in a home environment, interacting physically with human users and the surrounding environment. While the main target of RoboCup@Home is to evaluate the robots' abilities (e.g., the ability of gesture/speech recognition) and performance (e.g., tasks completed, errors made) in a realistic home environment setting, in the recent editions of the competition some tasks are designed exclusively to assess HRI aspects. Specifically, the HRI evaluation is performed in a peer-to-peer fashion: Any team involved in the study assigns a score to assess the HRI performance of the other teams.

Similarly to [24], we aim at quantifying the performance of social robots to execute HRI teaming tasks involving human users that actively participate in such tasks. However, in RoboCup@Home, the definition of HRI tasks is not driven by any dedicated methodology, and the conducted analysis neglects to investigating in detail the users' perception of robots' behavior, which is a crucial aspect to consider for assessing HRI in collaborative scenarios [56]. Even more important, the evaluation involves only expert users, whose previous knowledge of robots may seriously bias any potential finding. *Conversely, we approach these challenges through a methodology that clearly delineates the steps to design and evaluate HRI teaming tasks suitable to be executed in the range of a scientific competition with non-expert users. In addition, the evaluation is explicitly intended to explore the relationship between robots' performance and user perceptions, which is a novel aspect if compared with previous works that leverage robotic competitions as scientific experiments.*

3 METHODOLOGY

A robust HRI study performed in the range of a robotic competition requires careful planning and design [2]. As discussed in Section 2, many attempts were performed to leverage robotic competitions as experimental procedures to measure the performance of HRI tasks. However, the design and evaluation of such HRI tasks were always conducted in an ad-hoc way, without any clear methodological guidance. In this article, we tackle this issue by providing a general-purpose and repeatable methodology to support the design and evaluation of HRI teaming tasks in robotic competitions, tailoring the experiments to the specific competition at hand. Our methodology consists of the following steps:

(1) **Outline the characteristics of the selected robotic competition.** First, it is required to describe the context, objectives, scenarios, and overall vision of the robotic competition selected for the design of the HRI teaming task to be executed and evaluated. Note that designing a robotic competition from scratch is out of the scope of this article. Interested readers can refer to [2] to look at a set of guidelines to realize novel robotic competitions.

(2) **Design an HRI teaming task.** An HRI teaming task involves individuals, robots, objects, and courses of events referring to a collaborative scenario [60]. To provide an appropriate design of an HRI teaming task for social robots in a robotic competition, we rely on two relevant considerations delineated by previous HRI works: (i) HRI tasks should be designed to make them suitable to be executed by a robot [30]; (ii) HRI teaming tasks should be described through explicit scenarios that clarify the HRIs that may happen, thus inducing proper feedback on robots [61]. Based on the foregoing, and leveraging the framework discussed in [3] for HRI tasks' analysis in social contexts, we propose to characterize the design an HRI teaming task in a robotic competition using the following ingredients:

- **Name**, **Objectives** and **Duration** of the HRI teaming task to be designed.
- **Context** of the HRI task, presented as a short description of the task together with the list of conditions/constraints characterizing the situation in which the task should be done.
- **Users** involved and specification of their **Role** within the HRI task.
- **Teams** involved in the robotic competition and **Robots** employed to execute the HRI teaming task. The selection of robots for the task must be congruous with the domain being investigated [9].
- Details of the **Environment** in which the HRI task will be executed.
- Description of the **Scenario** underlying the HRI teaming task, i.e., the steps the robot(s) has(have) to take to achieve the objectives within the environment, making clear the role of the involved human participants.

(3) **Evaluate an HRI teaming task**. While the experience of interacting with a robot has been already proved to involve a strong social and emotional component [11], few researchers have directly explored how this affects the evaluation of the collaborative interaction between people and robots for executing an HRI teaming task [64]. In this direction, we contend that the evaluation of an HRI teaming task in the context of a robotic competition should cover the following methodological steps:

- **Define the Experimental Hypotheses:** The first step consists of determining the experimental hypotheses to be confirmed or the relationships to be explored through the execution of the HRI task of interest. A hypothesis is a statement about the relationship between two or more variables. It is a testable prediction about what is expected to happen in a study. In a *confirmatory research study*, there is a already a clear idea about the relationship between the variables under investigation, and the target is to investigate if a hypothesis is supported by data [5]. On the other hand, in an *exploratory research study*, there are no prior assumptions or hypotheses, and the aim is to uncover possible relationships between variables.
- **Determine the Study Design and Number of Users:** From the experimental hypotheses, it is possible to determine if the study design should be *within-subjects*, *between subjects,* or a *mixed-model factorial approach.* Of course, any of these approaches has strengths and weaknesses. In a within-subjects study, each user tests all the conditions being investigated. This reduces the error variance but decreases the quality of users' responses to the study due to repetitive presentations of the same task [46]. Conversely, in a between-subject study, users experience only one of the experimental conditions. A different group of users must be identified for each experimental condition to verify, and one single participant can be classified in only one of these groups. The results between the groups are then compared, without being biased by any "practice effect". However, results may be strongly impacted by the individual differences between the users of the different groups [41]. Finally, a mixed-model factorial design integrates both within-subjects and between-

subjects approaches, and it is useful for the exploration of the interaction effects between two or more independent variables. Nonetheless, the limitations of between-subject and within-subject studies both apply to the mixed-model factorial approach and must be considered. On the other hand, given that the adopted design choice will affect the definition of group size values, determining the appropriate number of users required to evaluate the experimental hypotheses in an HRI study can be still considered as a challenge in HRI [9]. To support the evaluators in this choice, the literature provides many power analysis tables that can be used to determine the (potentially) suitable amount of users to involve in the evaluation [51].

— **Select the Method(s) of Evaluation:** In the research literature, there exist four primary methods of evaluation for performing HRI studies: interviews and questionnaires [14], behavioral measures [33], psychophysiology measures [10], and task performance metrics [50]. Due to the fact that no one of them is without problems, often it is not sufficient to select the method that seems the most appropriate for the HRI task being investigated. According to [9], in order to obtain valid and reliable results to tackle the experimental hypotheses, it is important to consider using two or more methods of evaluation to gain a better understanding of HRI. While determining the best method(s) of evaluation for each possible HRI task at hand is out of the scope of this article (interested readers can refer to [9]), will show in Section 4 an effective questionnaire [56] specifically built for measuring the users' perception of robots' behavior in an HRI teaming task.

— **Conducting the Study and Evaluate the Results:** Once the experimental hypotheses to validate are defined, and the design study and method of evaluations are established, the designed HRI task can be finally executed according to the rules of the selected robotic competition. Applying the selected methods of evaluation, the objective is to collect data from the task executions to confirm/reject the experimental hypotheses [64].

4 DESIGNING AND EVALUATING AN HRI TASK WITHIN THE SCIROC COMPETITION

4.1 Outline the Characteristics of the Selected Robotic Competition

SciRoc is an EU-H2020 funded project supporting the **European Robotics League** (**ERL**) and whose purpose is to bring ERL tournaments in the context of smart cities, in order to show how robots will integrate in the cities of the future as physical agents living in them. The first SciRoc competition was held in the shopping mall of **Milton Keynes** (**MK**)[1] from 16 to 22 September 2019. The competition focused on smart shopping and was divided into a series of episodes, each consisting of a task to be performed through addressing specific research issues. Robots were required to cooperate with MK customers and with the ICT infrastructure of an MK shopping mall, 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, audio/visual inputs from CCTV cameras, crowd density sensor information, and many others.

In the range of SciRoc, robots were expected to execute tasks of different nature in three different situations: assisting customers, providing professional services, and supporting during emergency situations. Five episodes were finally selected by the SciRoc consortium (cf. Figure 1):

— *Delivery coffee shop order* (E3). The robot assist customers in a coffee shop to take care of customers, by taking orders and bringing objects to and from customers' tables.

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

Fig. 1. The five selected episodes for the first SciRoc competition: E3, E4, E7, E10, and E12.

— *Take the elevator* (E4). The robot takes an elevator crowded with customers to reach a service located in another floor.
— *Shipping pick and pack* (E7). The robot is located on one of the booths of the mall, and on the shelves, some goodies are displayed for sale to customers. Customers can place orders through a tablet. The robot collects the requested packages for the customer, placing them in a box.
— *Through the door* (E10). This episode is focused on opening and passing through a door.
— *Fast delivery of emergency pills* (E12). An aerial robot attends an emergency situation in which a first-aid kit needs to be delivered to a customer.

4.2 Design an HRI Teaming Task

Among the available episodes, we decided to focus on the most social one (E4) and to design a specific HRI teaming task following the methodology in Section 3.

— **Name:** Take the elevator (from now simply referred to as E4).
— **Objectives:** A robot must take an elevator of MK crowded with customers to reach another floor.
— **Duration:** Around 5–10 minutes.
— **Context:** 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. Note that when a floor is reached, the robot interacts (randomly) with one of the persons in the elevator asking her/him if the floor reached is the right one for it. Inside the elevator, the robot has to negotiate space and time of the person exiting. For example, if a floor is reached where a person has to go out, but not the robot. The robot should not block the door passage.
— **Users and their Role:** To create a realistic situation that can be encountered in a shopping mall, we decided to adopt the so-called "multi-heterogeneous humans to a single robot interaction" [39], involving four users randomly selected from the MK customers, each one assigned to a specific role. From two to three customers could have face-to-face interaction with the robot, while the other(s) was(were) observer(s) of the task. To be more specific, the robot encountered two persons while moving toward the elevator:

  — Person with **role A** was placed in a pre-defined location not far from the elevator. S/he could observe the robot but was not interested in interacting with it;
  — Person with **role B** actively moved towards the robot willing to interact with it, cf. Figure 2.

  Once arrived in front of the elevator, the robot encountered two further persons. Both persons took the elevator with the robot and were instructed to reach a specific floor, which could be different (or the same) from the one assigned to the robot. At this point, the robot interacted randomly with one of them asking to push (in place of it) the button for the floor it wants to reach. The two persons in the elevator played the following roles:



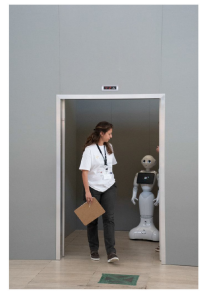

Fig. 2. Role A; Role B; Role C1/C2.

— Person with **role C1** always got off before the robot;
— Person with **role C2** got off together with (or after then) the robot, cf. Figure 2.

Last but not least, a user belonging to the SciRoc organizing team was selected as the referee for the task. The referee observed the execution of the task, ensuring a fair environment and identifying rule infractions.

— **Teams and Robots:** Five teams participated in the HRI task, employing different robot variants (see Figure 3):
  — **UC3M**[2] team, with researchers coming from the robotics laboratory of University Carlos III of Madrid. The team participated to the competition employing TIAGo,[3] a robot produced by PAL Robotics.[4] TIAGo is a mobile service robot originally designed to work in indoor environments. It has an extendable torso and its sensor suite allows it to perform a wide range of perception and navigation tasks. In addition to the basic robot's platform, the robot was also equipped by the team with a manipulator arm to grab tools and objects.
  — **Gentlebots**[5] team, with researchers in robotics from the Rey Juan Carlos University and the University of León. The team participated in the competition employing TIAGo robot. In addition to the basic robot's platform, the robot was also equipped with one tablet and one microphone in front, and the status of the robot was always shown on the tablet.
  — **HEARTS** team, with researchers coming from the Bristol Robotics Laboratory[6] that is focused on designing frameworks for developing assistive robots in the healthcare domain. The team participated in the competition employing Pepper[7] robot produced by Softbank.[8] Pepper is a human-like service robot that 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.
  — **eNTiTy** team of the R&D department of NTT Disruption in Spain. The researchers in the team focus on developing social robotics applications for clients. The team participated in the competition employing TIAGo. In addition to the basic robot's platform, the robot was also equipped by the team with a manipulator's arm and a signal light stuck on the head, which was able to change color according to the speech recognition status.

[2] https://github.com/roboticslab-uc3m.
[3] https://tiago.pal-robotics.com/.
[4] http://pal-robotics.com/.
[5] http://www.gentlebots.robotica.gsyc.es/.
[6] https://www.bristolroboticslab.com/.
[7] https://www.softbankrobotics.com/us/pepper.
[8] https://www.softbank.jp/en/robot/.

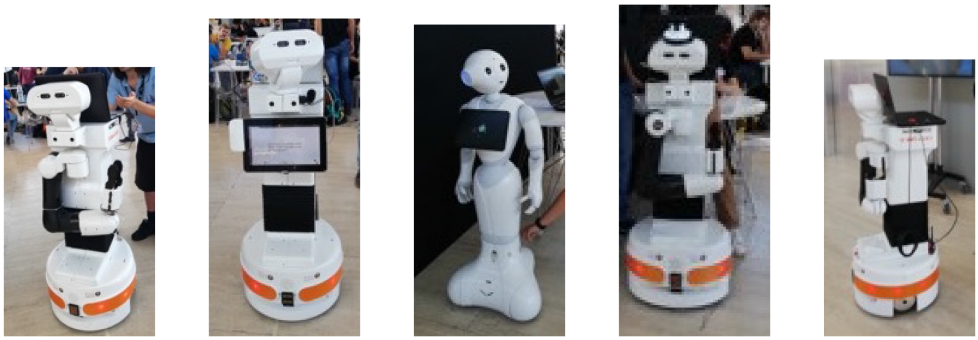

Fig. 3. Robots employed during E4, developed respectively by UC3M, Gentlebots, HEARTS, eNTITY, and LASR teams.

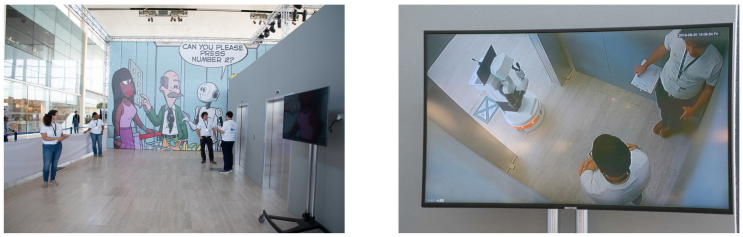

Fig. 4. Competition arena for the task E4.

- **LASR**[9] team, consists of AI researchers from the University of Leeds. The team participated in the competition employing TIAGo. In addition to the basic robot's platform, the robot was also equipped by the team with a manipulator's arm to grab tools and objects.

— **Details of the Environment:** The competition arena was set up in Milton Keynes Hall using a straight square truss system. Since the main target of the task was to evaluate HRI in a restricted space such as an elevator, to achieve a realistic dynamic social environment, we recreated a mock-up elevator inside the arena, complete with movable doors (see Figure 4). In addition, the elevator has been equipped with a video camera showing to the audience and to the experimenters the task in progress. The referee was physically positioned on the perimeter of the arena.

— **Description of the Task Scenario:** Based on the above elements, we designed the collaborative task scenario in the competition arena by sketching out its layout, the phases of the execution of the task, and the zones of the environment where the participants can move, see Figure 5.

— *Encounter Situation (Phase 1)*, the robot enters in the competition arena and continues the path toward zone A and zone B; role A and role B are deployed in either zone A or zone B randomly, and they can shift the zones during the run. The expected behavior of the robot is not to interact with role A and to interact with role B. More specifically, when the robot detects a participant who is not interested in interacting, the robot should just avoid her/him and proceed without any attempt to communicate. When the robot detects

[9] https://sensiblerobots.leeds.ac.uk/lasr/.

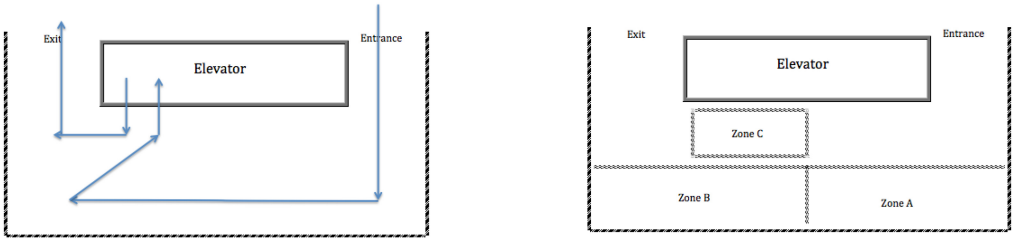

Fig. 5. The layout of the competition arena and zone division.

a person who is interested in interacting, the robot should stop moving and communicate with her/him accordingly.

— *Entering/exiting the elevator (Phase 2)*, once the robot passed zone B, it moves toward zone C. Zone C is occupied by two standing participants (roles C1/C2). The robot should place itself at a proper location outside the elevator depending on the location of the other participants. Elevators doors are operated by the referee after the robot signal that it has reached its desired location outside the elevator. When the elevator door opens, the robot has to wait until all the participants around the elevator enter the door,[10] then it can move and enter the door occupying a proper space in the elevator cabin. The robot is not able to press the button, hence, it has to face one participant, declare its target exit floor, and asks for help to press the button. The target exit floor of the robot has been communicated through MK data hubs. When the door opens, the referee declares the current floor, the robot is allowed to ask one participant (while facing her/him) which is the current floor. If the current floor is the target one for the robot, it must exit. Otherwise, it must stay and the elevator continues to "go up" to the next target floor. Each human participant in the task knows at the outset her/his target floor, and s/he should exit accordingly when the floor is reached. On the contrary, the participants' target exit floors are unknown to the robot. In case the robot has to exit with a participant, it must negotiate with the customer in the elevator who is going first. The door may open several times before reaching the robot's target exit floor.

— *Moving to exit (Phase 3)*, after exiting the elevator, the robot moves from the elevator to the finish area (exit).

4.3 Evaluate an HRI Teaming Task

— **Define the Experimental Hypotheses.** In the range of SciRoc, we decided to focus on investigating how human users involved in the selected HRI teaming task concretely perceive robots' behavior. In this direction, we devised three experimental hypotheses to be validated with a confirmatory research [5]:

— **[H1]** The robots' behavior perceived by users is influenced by the interaction modalities (e.g., voice tone, the complexity of the verbal communication, gestures employed) adopted by the robot.

— **[H2]** The robots' behavior perceived by users is influenced by users' gender.

— **[H3]** The robots' behavior perceived by users is influenced by users' role.

[H1] aims at confirming the impact of the robots' interaction modalities on the users' perception of the robots' behavior, even in the case in which the robots have slightly different

[10]In E4, we adopted the "polite behavior" for each robot involved in E4, which has to wait until all the users around the elevator enter the door before entering itself, according to the ethical guidelines made in [1].

Table 1. Schedule of the Runs

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Run 1	UC3M	eNTiTy	LASR	HEARTS	Gentlebots
Run 2	UC3M	eNTiTy	LASR	Gentlebots	HEARTS
Run 3	LASR	eNTiTy	UC3M	Gentlebots	HEARTS
Run 4	UC3M	Gentlebots	eNTiTy	LASR	HEARTS
Run 5	eNTiTy	LASR	UC3M	HEARTS	Gentlebots
Run 6	LASR	eNTiTy	UC3M	HEARTS	Gentlebots
Run 7	Gentlebots	HEARTS	UC3M	eNTiTy	LASR
Run 8	HEARTS	LASR	UC3M	Gentlebots	eNTiTy
Run 9	eNTiTy	HEARTS	LASR	UC3M	Gentlebots
Final Run	HEARTS	LASR	eNTiTy	Gentlebots	

appearances from each other (note that in SciRoc, four out of five teams used the same TIAGo robot, and the only difference lies in the robots' customization with ad-hoc software and hardware).

[H2] is on analyzing the impact of users' gender on users' perception in the context of the task. While there is already an evidence that males and females view robots differently, as already proved by many previous HRI research works, such as [12, 13, 15, 19, 26, 29, 35, 45, 48, 52, 55, 57], there are few results that explore how male and female users perceive the robot's behavior in case of an HRI teaming task performed in a robotic competition.

[H3] is about investigating the impact of users' role in the range of the task. The users' role has been already proved to be an important modulator of the perceived robot's behavior by the users [23, 25]. In SciRoc, we wanted to investigate which specific users' roles have had an impact on the perceived robots' behavior in E4.

Finally, we complemented the above study with an additional exploratory research aimed at analyzing the *relationship between the performance of robots in executing the selected HRI task and the users' perceptions of robots' behaviors* [H4].

— **Determine the Study Design and Number of Users.** The SciRoc competition lasted four days. In total, 10 runs of the HRI task were performed: nine runs in the first three days, and the final run took place on the last day of the competition. In the first nine runs, the five teams performed an execution of the task in a randomized order, according to the schedule shown in Table 1. The final run involved only the best four teams that performed better in the previous runs.

For conducting the evaluation, we decided to rely on a mixed-model factorial design. Specifically, the study involved a total of 40 users. The same four users participated only to one of the 10 runs. In any run, five different teams/robots (within-subject factor) performed the test according to the run schedule. User's gender (between-subject factor) was declared by users, before the starting of any run. Users' role (between-subject factor) was assigned by the task's referee (one user was assigned to role A, one user was assigned to Role B, two users were assigned to role C1/C2, respectively), before the starting of each run. Users and their roles were unknown to the robots.

— **Select the Method(s) of Evaluation.** We utilized two methods of evaluation (a questionnaire and a performance metric) to analyze the validity of the experimental hypotheses.
  — **Questionnaire.** At the end of any run, the participating users filled a dedicated questionnaire built ad-hoc for evaluating HRI teaming tasks [56]. The questionnaire has been

thought to specifically keep track of 17 behavioral aspects related to: *(i) social behavior of the robot*, *(ii) proxemics between human and robot*, and *(iii) collaboration with the robot*. The scores assigned on the scale range from: *Absolutely No = 1* to *Absolutely Yes = 5*. If compared with the original questionnaire [56], we decided to convert the only negative behavior, "Perceived Strangeness", into its "positive version", i.e., "Perceived Naturalness". The questions are organized as follows:

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 naturalness of the robot?

Proxemics between human and robot

* Did the robot look at your face during the conversation between user and the robot?
* Did you look at the robot's face during the conversation between user and the robot?
* Have you paid attention to the conversation with the robot?
* Have you 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?

While the validity of the adopted questionnaire is already discussed in our previous work [56], we further investigated its reliability by calculating the Cronbach's alpha coefficient (α) for the three macro-categories of the questionnaire. We obtained the following results: α of *Social Behavior of robot* = 0.907; α of *Proxemics between human and robot* = 0.921; α of *Collaboration with robot* = 0.83. According to [17], which discusses cut-off values for reliability indices, values of α coefficient greater than 0.8 indicate a reliability of the adopted scale among very good and excellent.

— **Task Performance.** We also assessed the robots' performance according to how well the teams/robots performed the HRI task execution, assigning them a score leveraging on a fair judging system. We employed two sets of scores, related to achievements and penalties. In addition, we determined the disqualifying behaviors according to the primary principles of HRI ethical [42] and social norms. If one of the disqualifying behaviors occurred, the performance was stopped and any score achieved so far was canceled.

Achievements

* The robot properly deals with the participant with role A (avoidance, no interaction).
* The robot properly deals with the participant with role B (interaction).
* The robot enters the elevator.
* The robot declares the target floor to the participant with role C1/C2.
* The robot exits the elevator at the proper floor.
* The robot reaches the finish area.

GENTLEBOTS			ENTITY			LASR			HEARTS			UC3M		
Run	A	P	Run	A	P	Run	A	P	Run	A	P	Run	A	P
1	3	1	1	1	0	1	1	2	1	0	0	1	0	0
2	1	0	2	6	0	2	1	0	2	0	0	2	0	0
3	5	0	3	3	2	3	2	0	3	1	0	3	0	0
4	4	0	4	5	0	4	3	0	4	1	0	4	0	0
5	1	0	5	5	0	5	1	0	5	5	0	5	3	0
6	5	0	6	0	0	6	6	0	6	1	0	6	1	0
7	5	0	7	1	0	7	2	0	7	3	0	7	1	0
8	4	0	8	5	0	8	6	0	8	5	0	8	2	1
9	5	0	9	4	0	9	1	0	9	1	0	9	2	0
Final1°-2°	4	0	Final 1°-2°	2	0	Final 3°-4°	7	0	Final3°-4°	5	0	-	-	-

Fig. 6. E4 Score Sheet.

* Above 80% of positive users' perceptions (>= 4) over total valid answers of the questionnaire.

Penalties

* Robot requires a participant to move away to avoid a collision.
* A participant instructs the robot to move away from one location.
* The robot acts participants' requests wrongly.
* The robot obstructs the way to the participants.
* Above 80% of negative users' perception (<= 2) over total valid answers of the questionnaire.

Disqualifying behaviors

* The robot hits a human.
* The robot hits and damages the furniture and/or objects.
* Team members give instructions to the robot during the task performance.

To encourage participating teams to better address HRI issues, as well as other robot's functionalities, we integrated the results of the user questionnaire analyzed in the previous point with achievements and penalties, with the target to reward teams with positive user evaluation, penalize teams with negative user evaluations, while keeping the score neutral for intermediate user evaluations.

5 RESULTS ANALYSIS

5.1 Evaluating Robots' Performance in E4

We evaluated the robots' performance in executing E4 according to the scores related to achievements and penalties (see Section 4). In each general run, the aggregate score has been determined by the third-highest score according to the ERL system. The top four teams in the ranking were qualified for the final. The final ranking for assigning the first, second and third place was determined by the performance in the Final. The E4 Score Sheet is shown in Figure 6.

As expected, the performance of the teams improved significantly throughout the SciRoc competition. The winner of E4 was the Gentlebots team, the eNTiTY team reached the second place, while the LARS team obtained the third place.

5.2 Analyzing the Results of the User Questionnaire

5.2.1 Data Collection. 40 users have participated in the 10 runs of E4. To ensure the heterogeneity of involved users, they were randomly selected from the MK customers. They were diversified

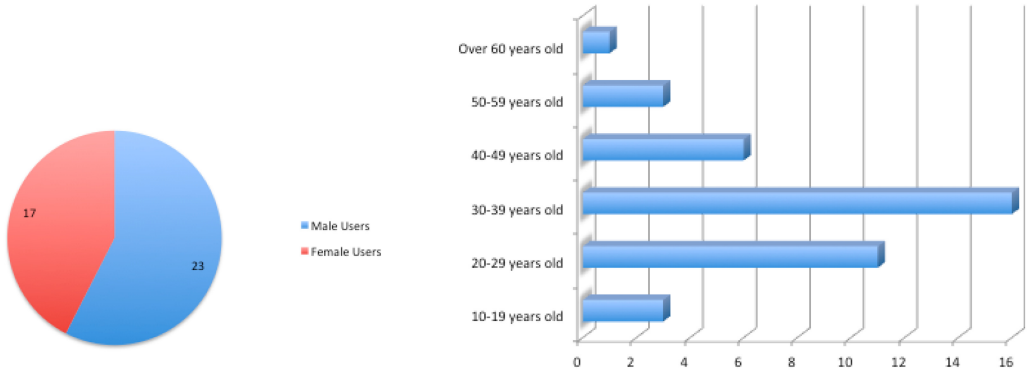

Fig. 7. Gender distribution and age distribution.

Table 2. Distribution of Users by Genders Per Run

		Male Users	Female Users
Day 1	Run 1	2	2
	Run 2	2	2
	Run 3	3	1
Day 2	Run 4	3	1
	Run 5	4	0
	Run 6	1	3
Day 3	Run 7	1	3
	Run 8	2	2
	Run 9	2	2
Day 4	Final Run	3	1

as secretaries, employees, students, retirees, and so on. Only one of them declared to have interacted with robots before. Gender distribution was 23 male and 17 female users. Age distribution was as follows: three users from 18–19 years old, 11 users from 20–29 years old, 16 users from 30–39 years old, six users from 40–49 years old, three users from 50–59 years old, one user over 60 years old (See Figure 7). The distribution of gender per run is shown in Table 2. We collected 196 questionnaires overall, of which 78 were considered incomplete (i.e., not filled at all or not filled completely because of a failed test in a run, which was a circumstance that happened especially on the first day of the competition). 118 questionnaires were considered completed.

5.2.2 Analyzing the Impact of Robots' Interaction Modalities on the Users' Perception. To validate the first experimental hypothesis [H1] (i.e., the robots' behavior perceived by users is influenced by the interaction modalities of the robots), we first completed the missing data using the Mean Imputation Method and then conducted Repeated Measures ANOVA to check how users perceived the robot behaviors among different developer teams. We found statistically significant differences in all the robot behaviors (i.e., p values of all the robot behaviors are less than 0.05). In the end, we conducted a Pairwise Comparison to check how users perceived differently between two participating teams. The results of significant values and mean difference values of pairwise comparison are shown in Tables 3, 4 and 5.

For the items related to *Social Behavior of the robot*, *Proxemics between human and robot*, *Collaboration with the robot*, we found that the robots employed by the UC3M team and Gentlebots

Table 3. Pairwise Comparison Between Participated Teams in Social Behavior

Social Behavior				
	Team (i)	Team (j)	Mean difference (i–j)	P-value
Perceived Happiness	eNTiTy	UC3M	0.897	**0.000**
	HEARTS	UC3M	0.897	**0.000**
	eNTiTy	LASR	0.573	0.029
	eNTiTy	Gentlebots	1,269	**0.000**
	HEARTS	LASR	0.573	0.014
	LASR	Gentlebots	0.688	0.010
	HEARTS	Gentlebots	1,260	**0.000**
Perceived Sociability	eNTiTy	UC3M	0.783	**0.001**
	HEARTS	UC3M	0.639	**0.000**
	eNTiTy	Gentlebots	1,210	**0.000**
	HEARTS	Gentlebots	1,066	**0.000**
Perceived Capability	eNTiTy	UC3M	0.463	0.041
Perceived Responsiveness	HEARTS	UC3M	0.417	0.035
	UC3M	Gentlebots	0.479	0.028
	eNTiTy	Gentlebots	0.813	**0.000**
	LASR	Gentlebots	0.631	0.044
	HEARTS	Gentlebots	0.896	**0.000**
Perceived Interactiveness	HEARTS	UC3M	0.480	0.016
	HEARTS	Gentlebots	0.733	**0.000**
***Perceived Naturalness**	eNTiTy	UC3M	0.306	0.037
	HEARTS	UC3M	0.835	**0.000**
	UC3M	Gentlebots	0.352	0.048
	HEARTS	eNTiTy	0.529	0.048
	eNTiTy	Gentlebots	0.658	**0.000**
	HEARTS	LASR	0.528	**0.007**
	LASR	Gentlebots	0.659	**0.007**
	HEARTS	Gentlebots	1,188	**0.000**

team were perceived as less sociable, less suitable in proxemics, and less collaborative than the others. While this result can not be attributed to the appearance of the robots (four out of the five teams involved in SciRoc employed exactly the same TIAGo robot to perform the HRI tasks), the influencing factor is related to the different interaction modalities adopted by the five robots to perform E4. It is worth noticing that the Gentlebots team was the winner in E4, meaning that the robot's performance alone does not represent the robots' behaviors perceived by users. We better investigate the relationship between robots' performance and user perception in [H4]. In the meanwhile, we can confirm the validity of H1 for the UC3M team and the Gentlebots team.

5.2.3 Analyzing the Impact of Users' Gender on the Perception of Robots' Behavior. To validate the second experimental hypothesis [H2] (i.e., the robots' behavior perceived by the users is influenced by users' gender), we first completed the missing data using Mean Imputation Method and then conducted Mixed-ANOVA to check how male and female users perceived differently the robot behavior. We found no *interaction effect* between with-in subject factor (i.e., teams) and the between-subject factor (i.e., gender), meaning that the impact of the between-subject factor does not depend on the level of with-in subject factor. However, we found highly significant difference

Table 4. Pairwise Comparison Between Participated Teams in Proxemics

Proxemics between human and robot				
	Team (i)	Team (j)	Mean difference (i–j)	P-value
Did robot look at your face ...	eNTiTy	UC3M	0.900	**0.000**
	LASR	UC3M	0.550	**0.003**
	HEARTS	UC3M	0.792	**0.000**
	eNTiTy	Gentlebots	1,160	**0.000**
	LASR	Gentlebots	0.810	**0.003**
	Hearts	Gentlebots	1,052	**0.000**
Did you look at the robot's face ...	eNTiTy	UC3M	0.992	**0.000**
	LASR	UC3M	0.900	**0.000**
	HEARTS	UC3M	0.700	**0.000**
	Gentlebots	UC3M	0.658	**0.000**
Have you paid attention to con ...	eNTiTy	UC3M	0.633	**0.000**
	LASR	UC3M	0.400	**0.000**
	HEARTS	UC3M	0.371	**0.005**
	eNTiTy	Gentlebots	0.500	**0.003**
Have you understood well con ...	eNTiTy	UC3M	0.538	**0.001**
	UC3M	Gentlebots	0.409	0.034
	eNTiTy	HEARTS	0.468	**0.001**
	eNTiTy	Gentlebots	0.947	**0.000**
	LASR	Gentlebots	0.758	**0.001**
	HEARTS	Gentlebots	0.479	0.011
Perceived Consciousness	eNTiTy	UC3M	0.639	**0.000**
	LASR	UC3M	0.447	0.018
	HEARTS	UC3M	0.792	**0.000**
	eNTiTy	Gentlebots	0.660	**0.001**
	HEARTS	Gentlebots	0.813	**0.000**
Perceived Friendliness	eNTiTy	UC3M	1,454	**0.000**
	LASR	UC3M	1,000	**0.000**
	HEARTS	UC3M	1,450	**0.000**
	eNTiTy	Gentlebots	1,575	**0.000**
	LASR	Gentlebots	1,120	**0.000**
	HEARTS	LASR	0.450	0.050
	HEARTS	Gentlebots	1,570	**0.000**
Perceived Politeness	eNTiTy	UC3M	1,050	**0.000**
	HEARTS	UC3M	1,042	**0.000**
	UC3M	Gentlebots	0.548	**0.001**
	eNTiTy	LASR	0.556	0.019
	eNTiTy	Gentlebots	1,598	**0.000**
	HEARTS	LASR	0.548	0.010
	LASR	Gentlebots	1,041	**0.000**
	HEARTS	Gentlebots	1,589	**0.000**
Perceived Adaptability	eNTiTy	UC3M	0.800	**0.000**
	LASR	UC3M	0.551	**0.002**
	HEARTS	UC3M	0.867	**0.000**
	eNTiTy	Gentlebots	0.598	0.011
	HEARTS	Gentlebots	0.664	**0.000**
Perceived Ease of use	eNTiTy	UC3M	0.696	**0.000**
	LASR	UC3M	0.603	**0.000**
	HEARTS	UC3M	0.655	**0.000**
	eNTiTy	Gentlebots	0.706	**0.001**
	LASR	Gentlebots	0.612	0.020
	HEARTS	Gentlebots	0.664	**0.001**

Table 5. Pairwise Comparison Between Participated Teams in Collaboration

Collaboration with robot				
	Team (i)	Team (j)	Mean difference (i–j)	P-value
Perceived Enjoyment	eNTiTy	UC3M	1,000	**0.000**
	LASR	UC3M	0.879	**0.000**
	HEARTS	UC3M	1,262	**0.000**
	eNTiTy	Gentlebots	0.845	**0.000**
	LASR	Gentlebots	0.724	**0.002**
	HEARTS	Gentlebots	1,107	**0.000**
Perceived Collaborativeness	eNTiTy	UC3M	1,038	**0.000**
	LASR	UC3M	0.990	**0.000**
	HEARTS	UC3M	1,089	**0.000**
	eNTiTy	Gentlebots	0.768	**0.000**
	LASR	Gentlebots	0.720	**0.002**
	HEARTS	Gentlebots	0.819	**0.000**

of *main effect* among the within subject factor, meaning that the overall effect over with-in subject effects is statistically significant.

For the items related to *Social Behavior* of the robot: *Perceived Responsiveness (p = 0.02)*, *Perceived Interactiveness (p = 0.03),* and *Perceived Naturalness (p = 0.019)*, we found significant differences between female and male users, meaning that female users perceived the robot's behavior more positively than male users. No other significant difference of between factor has been found in this analysis study. As a consequence, we can partially confirm the validity of our hypothesis. Only the social behaviors of robots, i.e., *Perceived Responsiveness*, *Perceived Interactiveness*, and *Perceived Naturalness* are influenced by users' gender.

Hence, we can confirm the findings of numerous research studies that gender differences affect the alteration of attitudes toward robots [37] and the perception of attitudes of robots [57]. Moreover, this finding could provide valuable references for designers and manufacturers of robots. For example, designers of social robots should make sure that the interaction style of the robot fits the users' gender and the users' individual attributes.

5.2.4 Analyzing the Impact of Users' Role on the Perception of Robots' Behavior. To validate the third experimental hypothesis [H3] (i.e., the robots' behavior perceived by the users is influenced by users' role), we first completed the missing data using Mean Imputation Method and then conducted Mixed-ANOVA to check how role A, role B, and role C1/C2 perceived differently the robots' behavior. We found no *interaction effect* between with-in subject factor (i.e., teams) and between-subject factor (i.e., role), meaning that the impact of between-subject factor does not depend on the level of with-in subject factor. However, we found a highly significant difference of *main effect* among the with-in subject factor.

For the items related to *Proxemics between human and robot*: *Have you paid attention to the conversation with the robot?* and *Have you understood well the meaning of conversation?*, we found significant differences between roles. Furthermore, after having conducted pairwise comparisons to check the effect among role A, role B and role C1/C2, we found remarkable significant difference between Role C1/C2 and Role A in *Have you paid attention to the conversation with the robot?* ($p = 0.008$), significant difference between role C1/C2 and role A in *Have you understood well the meaning of conversation?* ($p = 0.039$) and significant difference between role C1/C2 and role B in *Have you understood well the meaning of conversation?* ($p = 0.016$).

The results can be interpreted by concluding that in E4 the *users' role* has an impact on the users' perception of robots' behavior only when the interaction with the robots strongly involves spoken language or dialogues. As a consequence, we can partially confirm the validity of our hypothesis [H3].

5.2.5 Analyzing the Relationship between Robots' Performance and Users' Perception on Robots' Behaviors. The relationship between robots' performance and users' perception on robots' behaviors [H4] is one of the crucial issues that must be addressed in HRI. In E4, we were particularly interested in exploring if robots' behaviors perceived by users could be predictors of the performance scores. Since in E4 participating teams were penalized a few times and never disqualified, we approximated the achievements score as the overall score to investigate our exploratory research statement.

We first subtracted the scores related to "Above 80% of positive users' perceptions (> = 4) over total valid answers of questionnaire" from the achievements scores, and then we conducted a Multi Linear Regression study. We calculated R^2 value of regression for all the three macro-categories of the questionnaire, and obtained the following results: R^2 of Social Behavior of robot = 0.572; R^2 of Proxemics between human and robot = 0.431; R^2 of Collaboration with robot = 0.515. According to [40], R^2 represents the goodness of fit the model, whose cut-off value is 0,5. Hence, the macro category *Proxemics between human and robots* in the questionnaire cannot be further analyzed in this study.

For the items related to *Perceived Interactiveness* and *Perceived Collaborativenss*, we found significant values as follows: *Perceived Interactiveness* (p = 0.03 B = 0.78 t-value = 3.33) and *Perceived Collaborativenss* (p = 0.00 B = 0.842 t-value = 4,402), meaning those robots' behaviors can be considered as significant predictors of robots' performance score. The beta coefficient is the degree of change in the outcome variable for every 1-unit of change in the predictor variable. It can be positive or negative. In our case, we obtained two positive predictors: *Perceived Interactiveness* and *Perceived Collaborativenss*, i.e., for each single unit of positive change in users' perception of the interactivity of the robots' behavior, the performance score will increase by 0.78 as the degree of change; for each single unit of positive change in users' perception on collaborativeness of robot's behavior, performance score will increase 0.842 by the beta value. These results support the finding that there is a relationship between the performance score of E4 and the results of the questionnaire, i.e., the subjective evaluation of HRI by participants can indeed be reasonably approximated based on objective scoring. It is worth noticing that similar findings have been revealed in the context of VR in HRI [34].

5.3 Discussion on External Influencing Factors

From the point of view of human-related factors, *age effect,* and *cultural or ethnic effect* are often studied in HRI.

Users' age effect between different age groups may be significant; in particular, children and elders are susceptible to the impact of users' perception of HRI [36, 47, 54]. For example, in [19], the authors report that older participants are less willing to use the robot than younger ones in an experiment conducted by Robocare robot. In [49], the authors emphasize the importance to seek mutual gaze and switch addressee often in conversational robot for children. Since the groups of children and elders are not the target population in our research study, we have considered *users' age effect* as an external factor that may not influence the outcomes externally.

Users' cultural effect exists in both positive and negative attitudes towards robots [53, 58]. For instance, Li et al. [28] conclude that the cultural background predicts people's positive attitudes towards social robots: people from countries that have high exposure to industrial robots may have

less positive attitudes towards social robots. In [6, 8], the authors report that American users are the less negative towards robots, while Mexicans are the most negative, and Japanese participants do not show a particularly positive attitude towards robots. Furthermore, Lee and Šabanović [27] suggest that culturally variable attitudes and preferences towards robots are not simply reducible to factors such as perceptions and acceptances, rather they relate to more specific social dynamics and norms. In our case, the SciRoc organization committee emphasizes the importance of the principle of diversity when they selected the target population. The users participating to the task came from different backgrounds, the users were diversified as Asian, European, African, Muslims, Christians, Buddhists, and so on. The users were selected by the SciRoc organization committee randomly from the target population. In our research study, *users' cultural effect* is an uncontrollable factor that we believe can not influence externally robots' behaviors perceived by users in the case of E4.

From the point of view of robot-related factors, we analyzed robots' performance as an internal influencing factor of robots' behaviors perceived by users. However, a social robot is a manifestation of the human characteristics and human actions. Hence, robots' appearance, i.e., *anthropomorphism of robot*, may also influence the outcomes externally in E4. According to Fong et al. [16], we can classify robots based on their appearance into four categories: anthropomorphic, zoomorphic, caricatured, and functional. People behave differently when interacting with a pet robot and with a humanoid robot [4]. Robots that are human-like in both appearance and behavior are treated less harshly than machine-like robots [7]. In the field of social robots, there is an increased tendency to build robots that resemble humans in their appearance. The five robots that participated in E4 were all wheeled human-like robots: one Pepper robot (social robots) and four TIAGo robots (service robots). Teams could slightly modify their robots' appearance (see in Figure 3). Since all robots were humanoid robots in E4, we have considered *anthropomorphism of robot* as an external factor that may not influence the robots' behaviors perceived by users.

6 CONCLUDING REMARKS

The experience of interacting with a robot has been shown to be very different in comparison to people's interaction experience with other technologies and artifacts, and has been proved to have a strong social or emotional component'a difference that poses potential challenges related to the design and evaluation of HRI.

In this article, we have addressed this issue by presenting a general-purpose and repeatable methodology for conducting studies in collaborative HRI in the range of robotic competitions. The methodology includes a step-by-step approach to design HRI teaming tasks tailored to be enacted in a robotic competition and to evaluate the performance of social robots to execute the designed tasks, exploring the relationship between robots' performance and user perceptions based on the feedback of the users participating to such tasks. We have assessed the feasibility of the methodology by instantiating it over a real robotic competition SciRoc, to show its feasibility to design and evaluate a concrete HRI task.

The focus of SciRoc is the interaction among humans, autonomous robots, and smart cities, or more in general, to showcase to the general public how robots can coexist in a public scenario. Involving external non-experts users in robot competitions can promote the dissemination of HRI scientific research, and improve the visibility of AI and robotics technologies for audiences. In this direction, the proposed methodology can be considered as a relevant achievement for designing and evaluating HRI teaming tasks based on robotic competitions.

The evaluation conducted on one task specifically designed with our methodology in the context of SciRoc has enabled us to obtain interesting findings indicating that: (i) even in the case of robots having an almost identical appearance, slightly changing the interaction modalities can affect how

the users' perceive the robots' behavior; (ii) only some social behaviors of robots are concretely influenced by the users' gender; (iii) the users' role has an impact on the users' perception of robots' behavior only when the interaction with the robots is conducted in spoken language; (iv) robots' behaviors perceived by users can be (sometimes) predictors of robots' performance in executing an HRI teaming task, in particular for the behaviors related to *Perceived Interactiveness* and *Perceived Collaborativeness*. Moreover, we got an interesting finding about the *Gentlebots* team, the winner of the competition, which got worse on users' perceptions (i.e., results of questionnaire) compared with the other teams, meaning that the competition scoring should emphasize more on users' perceptions of the robot rather than on robots' functionalities.

It is worth noting that the results obtained from the application of our methodology to SciRoc represent alone (i.e., independently by the methodology) interesting findings in the range of HRI research. Nonetheless, as a future work, we aim at using our methodology to design and evaluate further HRI tasks in other robotic competitions than SciRoc, with the aim at verifying if the inferred findings can be generalized also in other applications scenarios.

REFERENCES

[1] Sabah S. Al-Fedaghi. 2008. Typification-based ethics for artificial agents. In *Proceedings of the 2008 2nd IEEE International Conference on Digital Ecosystems and Technologies*. IEEE, 482–491.

[2] Francesco Amigoni, Emanuele Bastianelli, Jakob Berghofer, Andrea Bonarini, Giulio Fontana, Nico Hochgeschwender, Luca Iocchi, Gerhard Kraetzschmar, Pedro Lima, Matteo Matteucci, Pedro Miraldo, Daniele Nardi, and Viola Schiaffonati. 2015. Competitions for benchmarking: Task and functionality scoring complete performance assessment. *IEEE Robotics & Automation Magazine* 22, 3 (2015), 53–61.

[3] Elisabeth André, Ana Paiva, Julie Shah, and Selma Šabanovic. 2020. Social agents for teamwork and group interactions (dagstuhl seminar 19411). In *Proceedings of the Dagstuhl Reports*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[4] Anja Austermann, Seiji Yamada, Kotaro Funakoshi, and Mikio Nakano. 2010. How do users interact with a pet-robot and a humanoid. In *Proceedings of the CHI'10 Extended Abstracts on Human Factors in Computing Systems*. 3727–3732.

[5] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-robot Interaction: An Introduction*. Cambridge University Press.

[6] Christoph Bartneck, Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kennsuke Kato. 2005. Cultural differences in attitudes towards robots. In *Proceedings of the SSAISB 2005 Convention on Social Intelligence and Interaction in Animals, Robots and Agents*. 1–4.

[7] Christoph Bartneck, Juliane Reichenbach, and Julie Carpenter. 2006. Use of praise and punishment in human-robot collaborative teams. In *Proceedings of the ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 177–182.

[8] Christoph Bartneck, Tomohiro Suzuki, Takayuki Kanda, and Tatsuya Nomura. 2007. The influence of people's culture and prior experiences with aibo on their attitude towards robots. *Ai & Society* 21, 1–2 (2007), 217–230.

[9] Cindy L. Bethel and Robin R. Murphy. 2010. Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics* 2, 4 (2010), 347–359.

[10] Cindy L. Bethel, Kristen Salomon, Robin R. Murphy, and Jennifer L. Burke. 2007. Survey of psychophysiology measurements applied to human-robot interaction. In *Proceedings of the RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 732–737.

[11] Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International Journal of Human-computer Studies* 59, 1–2 (2003), 119–155.

[12] Meia Chita-Tegmark, Monika Lohani, and Matthias Scheutz. 2019. Gender effects in perceptions of robots and humans with varying emotional intelligence. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 230–238.

[13] Kerstin Dautenhahn, Michael Walters, Sarah Woods, Kheng Lee Koay, Chrystopher L. Nehaniv, A. Sisbot, Rachid Alami, and Thierry Siméon. 2006. How may I serve you? A robot companion approaching a seated person in a helping context. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*. 172–179.

[14] Maartje De Graaf, Somaya Ben Allouch, and Jan Van Diik. 2017. Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study. In *Proceedings of the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 224–233.

[15] Friederike Eyssel, Dieta Kuchenbrandt, Frank Hegel, and Laura de Ruiter. 2012. Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *Proceedings of the 2012 IEEE RO-MAN: The 21st*

IEEE International Symposium on Robot and Human Interactive Communication. IEEE, 851–857.
[16] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and Autonomous Systems* 42, 3–4 (2003), 143–166.
[17] S. George and L. Mallery. 2003. Alfa de cronbach y consistencia interna de los ítems de un instrumento de medida. *Revista de Estudios Interdisciplinarios en Ciencias Sociales* 3, 16 (2003), 3–9.
[18] Michael A. Goodrich and Alan C. Schultz. 2008. *Human-robot Interaction: A Survey*. Now Publishers Inc.
[19] Marcel Heerink. 2011. Exploring the influence of age, gender, education and computer experience on robot acceptance by older adults. In *Proceedings of the 2011 6th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 147–148.
[20] Anna Henschel, Ruud Hortensius, and Emily S. Cross. 2020. Social cognition in the age of human-robot interaction. *Trends in Neurosciences* 43, 6 (2020), 373–384. DOI: https://doi.org/10.1016/j.tins.2020.03.013
[21] Shah Rukh Humayoun, Tiziana Catarci, Massimiliano de Leoni, Andrea Marrella, Massimo Mecella, Manfred Bortenschlager, and Renate Steinmann. 2009. Designing mobile systems in highly dynamic scenarios: The WORKPAD methodology. *Knowledge, Technology & Policy* 22, 1 (2009), 25–43.
[22] Shah Rukh Humayoun, Tiziana Catarci, Massimiliano de Leoni, Andrea Marrella, Massimo Mecella, Manfred Bortenschlager, and Renate Steinmann. 2009. The WORKPAD user interface and methodology: Developing smart and effective mobile applications for emergency operators. In *Proceedings of the International Conference on Universal Access in Human-Computer Interaction*. Springer, 343–352.
[23] Helge Hüttenrauch and Kerstin Severinson-Eklundh. 2006. To help or not to help a service robot: Bystander intervention as a resource in human–robot collaboration. *Interaction Studies* 7, 3 (2006), 455–477.
[24] Luca Iocchi, Dirk Holz, Javier Ruiz-del Solar, Komei Sugiura, and Tijn Van Der Zant. 2015. RoboCup@ Home: Analysis and results of evolving competitions for domestic and service robots. *Artificial Intelligence* 229 (2015), 258–281.
[25] Theodora Koulouri, Stanislao Lauria, Robert D. Macredie, and Sherry Chen. 2012. Are we there yet? The role of gender on the effectiveness and efficiency of user-robot communication in navigational tasks. *ACM Transactions on Computer-Human Interaction* 19, 1 (2012), 1–29.
[26] I. Han Kuo, Joel Marcus Rabindran, Elizabeth Broadbent, Yong In Lee, Ngaire Kerse, RMQ Stafford, and Bruce A. MacDonald. 2009. Age and gender factors in user acceptance of healthcare robots. In *Proceedings of the RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 214–219.
[27] Hee Rin Lee and Selma Šabanović. 2014. Culturally variable preferences for robot design and use in south korea, turkey, and the united states. In *Proceedings of the 2014 9th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 17–24.
[28] Dingjun Li, P. L. Patrick Rau, and Ye Li. 2010. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics* 2, 2 (2010), 175–186.
[29] Chun Hung Lin, Eric Zhi Feng Liu, and Yuan Yen Huang. 2012. Exploring parents' perceptions towards educational robots: Gender and socio-economic differences. *British Journal of Educational Technology* 43, 1 (2012), E31–E34.
[30] Norjasweeen Abdul Malik, Hanafiah Yussof, Fazah Akhtar Hanapiah, Rabiatul Adawiah Abdul Rahman, and Husna Hassan Basri. 2015. Human-robot interaction for children with cerebral palsy: Reflection and suggestion for interactive scenario design. *Procedia Computer Science* 76 (2015), 388–393.
[31] Andrea Marrella, Massimo Mecella, and Alessandro Russo. 2011. Collaboration on-the-field: Suggestions and beyond. In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management*.
[32] Jeremy A. Marvel, Shelly Bagchi, Megan Zimmerman, Murat Aksu, Brian Antonishek, Yue Wang, Ross Mead, Terry Fong, and Heni Ben Amor. 2020. Test methods and metrics for effective HRI in real-world human-robot teams. In *Proceedings of the Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 652–653.
[33] Amandine Mayima, Aurélie Clodic, and Rachid Alami. 2021. Towards robots able to measure in real-time the quality of interaction in HRI contexts. *International Journal of Social Robotics* 14, 3 (2021), 1–19.
[34] Yoshiaki Mizuchi and Tetsunari Inamura. 2020. Optimization of criterion for objective evaluation of HRI performance that approximates subjective evaluation: A case study in robot competition. *Advanced Robotics* 34, 3–4 (2020), 142–156.
[35] Bilge Mutlu, Steven Osman, Jodi Forlizzi, Jessica Hodgins, and Sara Kiesler. 2006. Task structure and user attributes as elements of human-robot interaction design. In *Proceedings of the ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 74–79.
[36] Tatsuya Nomura and Akira Nakao. 2010. Comparison on identification of affective body motions by robots between elder people and university students: A case study in japan. *International Journal of Social Robotics* 2, 2 (2010), 147–157.
[37] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Altered attitudes of people toward robots: Investigation through the negative attitudes toward robots scale. In *Proceedings of the AAAI-06 Workshop on Human Implications of Human-robot Interaction*. 29–35.
[38] Adam Norton, Willard Ober, Lisa Baraniecki, Eric McCann, Jean Scholtz, David Shane, Anna Skinner, Robert Watson, and Holly Yanco. 2017. Analysis of human–robot interaction at the DARPA robotics challenge finals. In *Proceedings*

of the International Journal of Robotics Research 36, 5–7 (2017), 483–513.
[39] Priyam Parashar, Lindsay M. Sanneman, Julie A. Shah, and Henrik I. Christensen. 2019. A taxonomy for characterizing modes of interactions in goal-driven, human-robot teams. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2213–2220.
[40] Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. 1973. *Linear Statistical Inference and Its Applications*. Wiley New York.
[41] Christian Remy, Oliver Bates, Jennifer Mankoff, and Adrian Friday. 2018. Evaluating HCI research beyond usability. In *Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–4.
[42] Laurel Riek and Don Howard. 2014. A code of ethics for the human-robot interaction profession. *Proceedings of We Robot* (2014). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757805.
[43] Laurel D. Riek. 2012. Wizard of oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136. DOI: https://doi.org/10.5898/JHRI.1.1.Riek
[44] Silvia Rossi, Francois Ferland, and Adriana Tapus. 2017. User profiling and behavioral adaptation for HRI: A survey. *Pattern Recognition Letters* 99 (2017), 3–12.
[45] Paul Schermerhorn, Matthias Scheutz, and Charles R. Crowell. 2008. Robot social presence and gender: Do females view robots differently than males? In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*. 263–270.
[46] Wendy A. Schweigert. 2021. *Research Methods in Psychology: A Handbook*. Waveland Press.
[47] Suleman Shahid, Emiel Krahmer, Marc Swerts, and Omar Mubin. 2010. Child-robot interaction during collaborative game play: Effects of age and gender on emotion and experience. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*. 332–335.
[48] Rosanne M. Siino and Pamela J. Hinds. 2005. Robots, gender & amp; sensemaking: Sex segregation's impact on workers making sense of a mobile autonomous robot. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2773–2778.
[49] Gabriel. Skantze. 2017. Predicting and regulating participation equality in human-robot conversations: Effects of age and gender. In *Proceedings of the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 196–204.
[50] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*. 33–40.
[51] James P. Stevens. 2013. *Intermediate Statistics: A Modern Approach*. Routledge.
[52] Megan Strait, Priscilla Briggs, and Matthias Scheutz. 2015. Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. In *Proceedings of the 4th international Symposium on New Frontiers in Human Robot Interaction*. 21–22.
[53] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2009. The negative attitudes to wards robots scale and reactions to robot behaviour in a live human-robot interaction study. In *Proceedings of AISB09 - 23rd Convention of the Society for the Adaptive and Emergent Behaviour and Complex Systems*.
[54] Fang-Wu Tung. 2011. Influence of gender and age on the attitudes of children towards humanoid robots. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 637–646.
[55] M. Walters, Samuel Marcos, Dag Sverre Syrdal, and Kerstin Dautenhahn. 2013. An interactive game with a robot: People's perceptions of robot faces and a gesture-based user interface. In *Proceedings of the 6th International Conference on Advances in Computer–Human Interactions*. 123–128.
[56] Lun Wang, Luca Iocchi, Andrea Marrella, and Daniele Nardi. 2019. Developing a questionnaire to evaluate customers' perception in the smart city robotic challenge. In *Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 1–6.
[57] Lun Wang, Andrea Marrella, and Daniele Nardi. 2019. Investigating user perceptions of HRI in social contexts. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 544–545.
[58] Astrid Weiss, Regina Bernhaupt, Manfred Tscheligi, and Eiichi Yoshida. 2009. Addressing user experience and societal impact in a user study with a humanoid robot. In *Proceedings of the AISB2009: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction*. Citeseer, 150–157.
[59] Astrid Weiss, Thomas Scherndl, Manfred Tscheligi, and Aude Billard. 2009. Evaluating the ICRA 2008 HRI challenge. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*. 261–262.
[60] Peter Wright. 1992. What's in a scenario? *ACM SIGCHI Bulletin* 24, 4 (1992), 11–12.
[61] Qianli Xu, Jamie Ng, Odelia Tan, Zhiyong Huang, Benedict Tay, and Taezoon Park. 2015. Methodological issues in scenario-based evaluation of human–robot interaction. *International Journal of Social Robotics* 7, 2 (2015), 279–291.
[62] Holly A. Yanco, Jill L . Drury, and Jean Scholtz. 2004. Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition. *Human–Computer Interaction* 19, 1–2 (2004), 117–149.

[63] Holly A. Yanco, Adam Norton, Willard Ober, David Shane, Anna Skinner, and Jack Vice. 2015. Analysis of human-robot interaction at the darpa robotics challenge trials. *Journal of Field Robotics* 32, 3 (2015), 420–444.
[64] James E. Young, JaYoung Sung, Amy Voida, Ehud Sharlin, Takeo Igarashi, Henrik I. Christensen, and Rebecca E. Grinter. 2011. Evaluating human-robot interaction. *International Journal of Social Robotics* 3, 1 (2011), 53–67.

Received October 2020; revised November 2021; accepted March 2022