Faculty of Engineering & Information Sciences

Proceedings of Metrics for Human-Robot Interaction

A workshop at the Third ACM/IEEE International Conference on Human-Robot Interaction (HRI08)

12 March 2008

Catherina R. Burghart

Aaron Steinfeld

Technical Report 471

School of Computer Science, University of Hertfordshire

February 2008

The workshop "Metrics for Human-Robot Interaction" is part of the Third ACM/IEEE International Conference on Human-Robot Interaction (HRI08), Amsterdam, The Netherlands, 12-15 March 2008

HRI08 Organizing Committee

General Chairs	Kerstin Dautenhahn *(University of Hertfordshire, UK)* Terry Fong *(NASA Ames Research Center, USA)*
Program Chairs	Matthias Scheutz *(Indiana University Bloomington, USA)* Yiannis Demiris *(Imperial College, England)*

Exhibitions Chairs	Holly Yanco *(University of Massachusetts Lowell, USA)* Christoph Bartneck *(Eindhoven University of Technology, The Netherlands)*
Finance Chairs	Julie A. Adams *(Vanderbilt University, USA)* Curtis Nielsen *(Idaho National Labs, USA)*
Local Arrangements	Ben Kröse *(University of Amsterdam, The Netherlands)* Marcel Heenrik *(Hogeschool van Amsterdam, The Netherlands)*
Publicity Chairs	Takayuki Kanda *(ATR, Japan)* Geb Thomas *(University of Iowa, USA)* Vanessa Evers *(University of Amsterdam, The Netherlands)*
Registration Chairs	Guido Bugmann *(University of Plymouth, UK)*
Workshops & Tutorials Chair	Kerstin Severinson Eklundh *(KTH, Sweden)*
Video Session Chair	Christoph Bartneck *(Technische Universiteit Eindhoven, The Netherlands)*

Metrics for Human-Robot Interaction 2008

A workshop affiliated with the 3rd ACM/IEEE International Conference on Human-Robot Interaction

March 12th, Felix Meritis, Amsterdam

Edited by:

Chair

Catherina R. Burghart
Institute of Process Control and Robotics
University of Karlsruhe
burghart@ira.uka.de

Co-Chair

Aaron Steinfeld
Robotics Institute
Carnegie Mellon University
steinfeld@cmu.edu

Presentation of the enclosed papers was preceded by an invited talk by Julie A. Adams (Vanderbilt University).

The workshop program committee included: Brian Scassellati (Yale University), Alan Schultz (Naval Research Laboratory), Chad (Odest) Jenkins (Brown University), Gal Kaminka (Bar Ilan University), and Ralf Mikut (KIT).

Table of Contents

Human-Robot Interaction Metrics and Future Directions
C. R. Burghart & A. Steinfeld 1

Teams & Frameworks

Steps to Creating Metrics for Human-like Movements and Communication Skills (of Robots)
H. Holzapfel, R. Mikut, C. R. Burghart, & R. Häußling 3

Identifying Generalizable Metric Classes to Evaluate Human-Robot Teams
P. Pina, M. L. Cummings, J. W. Crandall, & M. Della Penna 13

Toward Developing HRI Metrics For Teams: Pilot Testing In the Field
J. Burke, K. S. Pratt, R. Murphy, M. Lineberry, M. Taing, & B. Day 21

Framing and Evaluating Human-Robot Interactions
C. W. Nielsen, D. J. Bruemmer, D. A. Few, & D. I. Gertman 29

Social and Physical Interaction

Measuring the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots
C. Bartneck, D. Kulic, & E. Croft 37

Social Resonance: a Theoretical Framework and Benchmarks to Evaluate the Social Competence of Humanoid Robots
T. Chaminade 45

The Utility of Gaze in Spoken Human-Robot Interaction
M. Staudte & M. Crocker 53

A Visual Method for Robot Proxemics Measurements
T. van Oosterhout & A. Visser 61

Human-Robot Interaction Metrics and Future Directions

Catherina R. Burghart
Institute of Process Control and Robotics
University of Karlsruhe
burghart@ira.uka.de

Aaron Steinfeld
Robotics Institute
Carnegie Mellon University
steinfeld@cmu.edu

ABSTRACT
The Metrics for Human-Robot Interaction 2008 workshop at the 3rd ACM/IEEE International Conference on Human-Robot Interaction was initiated and organized to further discussion and community progress towards metrics for human-robot interaction (HRI). This report contains the papers presented at the workshop, background information on the workshop itself, and future directions underway within the community.

Categories and Subject Descriptors
I.2.9 [Artificial Intelligence]: Robotics – operator interfaces; H.5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces – Evaluation/methodology; J.4 [Computer Applications]: Social and Behavioral Sciences – Psychology.

General Terms
Measurement, Experimentation, Human Factors.

Keywords
Human-robot interaction, evaluation, metrics.

1. OVERVIEW
The evaluation of interactions between robots and humans, different types of interactions, and individual robot and human behaviors require adequate metrics and guidelines. These metrics should take into account a variety of factors, ranging from objective performance to social interaction. These metrics can code behaviors, ways of interaction, social and psychological aspects, and technical characteristics or objective measures (i.e. success rates, interaction time, error rates, etc.). There are metrics that can be acquired using objective measuring tools; others depend on the personal interpretation by the staff conducting and analysing experiments. When human beings are present, metrics for social human-robot interaction are of utmost interest in order to achieve robotic systems that can be intuitively handled by people without causing frustration and despair.

The following key questions were cited as topics of interest during the call for papers:

1. Which guidelines should be followed for careful experimentation in HRI?
2. Are there objective metrics applicable to HRI?
3. Are there social metrics applicable to HRI?
4. What is the relevance of subjective criteria for evaluation?
5. Can subjectively categorized criteria be used to form objective metrics for social HRI?
6. How can benchmarks (standardized tasks) be used to evaluate human-robot interactions?

2. WORKSHOP GOALS
The goals of the workshop was to propose guidelines for the analysis of human-robot experiments and forward a handbook of metrics that would be acceptable to the HRI community and allow researchers both to evaluate their own work and to better assess the progress of others. To achieve these goals the intended workshop format combined information about different metrics and evaluation methods given in submitted and invited talks, and moderated group discussions.

3. FUTURE DIRECTIONS

3.1 Workshop publications
Besides proceedings in this printed report, all presentation materials will be available on the workshop webpage (http://www.hri-metrics.org/metrics08). The aim of the workshop is to come up with a set of guidelines for experimental evaluation and a handbook of different metrics. It is intended to publish the results as well as selected papers in a special issue of an international journal.

3.2 Further community interaction
Under funding from the U. S. National Science Foundation (CBET-0742350), the website used for this workshop's call for papers (http://www.hri-metrics.org) will soon host a collaborative infrastructure for continued collaboration and discussion. The website will support evaluation documentation, data collection, data sharing, and cross-study comparisons.

This funding is also supporting integration of native data contribution to this community website in USARSim, an existing HRI simulation research tool (http://usarsim.sourceforge.net).

4. ACKNOWLEDGMENTS
We would like to thank Kerstin Dautenhahn and the University of Hertfordshire for their work towards the printing of this report.

Part of this work is based on work supported by the National Science Foundation under Grand No. CBET-0742350.

Steps to Creating Metrics for Human-like Movements and Communication Skills (of Robots) *

Hartwig Holzapfel
University of Karlsruhe,
Interactive Systems Lab.
D-76128 Karlsruhe, Germany
hartwig@ira.uka.de

Ralf Mikut
Forschungszentrum Karlsruhe
GmbH, Institute for Applied
Computer Science
P.O. Box 3640
76021 Karlsruhe, Germany
ralf.mikut@iai.fzk.de

Catherina Burghart
University of Karlsruhe,
Institute of Process Control
and Robotics
D-76128 Karlsruhe, Germany
burghart@ira.uka.de

Roger Häußling
University of Karlsruhe
D-76128 Karlsruhe, Germany
roger.haeussling@soziologie.uka.de

ABSTRACT

When assessing interactions between robots and persons existing metrics do mainly take into account aspects that are easy to measure like performance metrics; the human element is left out although it plays a central role in human-robot interaction. Measuring and quantifying human behavior as well as subjective impressions is rather complicated, may take a lot of time and ties a lot of labor. Means for quantifying human behavior usually cannot be automated; the analysis of video recordings, interviews and questionnaires requires many people who use their subjective judgment. By having several persons code or analyze the same data, subjective votes can be reduced. In this paper we present steps to build a framework of metrics for the evaluation of human-robot interaction focusing on communication and movement analysis. Besides using metrics to asses technical merits we present how the cooperation between engineers, computer scientists, and sociologists helps to code and quantify human behavior. Sociologists focus on interaction processes and structures like roles, communication rules, expectations, etc. Cognitive processes of the actors themselves (the interacting persons or robots) are not analyzed by sociologists; for this purpose psychologists are needed.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

General Terms

Evaluation metrics, Motion analysis, Language Processing

*Supported by DFG.

1. INTRODUCTION

Today, robots can be found in more and more environments of typical human everyday life i.e. in hospitals, hotels, museums, schools, and households. Thus robot encounters and robot interaction with naive persons are predestined. The usage of robotic devices in the human world requires appropriate design of the robot's interface to the environment and of the robot's cognitive skills, thus enabling intuitive interaction between robot and people. It is simply impossible for robot designers to anticipate human behavior and human interaction strategies; this often leads to interactions being exhausting, enervating or frustrating for the interacting person. However, the person's feeling of satisfaction with an interacting robot is difficult to measure.

Appropriate metrics defined to assess robot behavior in all fields of robotic life (technical merits, speech recognition and understanding, successful interaction, human and robot behavior, proximity, cognitive skills) can help to quantify and qualify interactions between robots and persons and to give essential feedback to robotic engineers. Whereas many international research groups focus on designing intelligent service robots, only some research has been done in the field of creating and employing objective and adequate criteria to evaluate human-robot interaction, so far.

One common means to assess robot success are benchmarks. A great variety of benchmarks do exist: some recent examples with a great deal of public attention are robot soccer competitions in different leagues [5], test parcours for rescue robots [19], the DARPA Grand Challenge [30] and the DARPA Urban Grand Challenge [1] for autonomous driving of cars. Human-robot interaction is contemplated by the RoboCup@Home league founded in 2007 [25]. The performance measurement is based on a score derived from competition rules and the evaluation by a jury. However, a transfer of such competition concepts and evaluation metrics to domains in the human everyday world can cover only a part of the necessary evaluation procedures.

Besides competitions, various metrics are used by interna-

tional researchers like the preferred direction of approaching in a living room scenario [35] or the distance a person feels most comfortable with when interacting with a robot [33]. Others, as proposed by [29] include success rates and number of operator interventions in tele-operated scenarios. Additionally, metrics for performance, world complexity and information quantification were established for autonomous mobile robots navigating in a corridor clotted by random obstacles [22]. In the first category instantaneous velocity, traveled distance, mission duration, mission success rate and power usage were measured, whereas global complexity and the vicinity of the robot are taken into account in the second category. The last metric used is the conditional entropy measuring the information contained in the internal robot map compared to the world map.

As soon as communication forms an integral part of human-robot interaction additional objective metrics like WER: word error rate (the standard metric for automatic speech recognition - ASR), CER: concept error rate (error rate to measure understanding, based on recognized concepts) and TER: turn error rate (based on number of turns that cannot be transformed to the correct semantics) can be applied. Current research on spoken dialog system uses either objective metrics, subjective metrics, or both. The main advantages of subjective metrics over objective metrics are that the user's subjective perception of the system can be included in the evaluation. Most measurements are based on questionnaires with rating scales such as Likert-Scales. Approaches exist to build a unified framework for the evaluation of dialog systems and create comparable scores with the PARADISE framework [32] for spoken dialog systems.

In contrast to metrics based on measurable characteristics and typically used in engineering, [20] suggest metrics for human-robot interaction devised from an psychologist's point of view which include autonomy, imitation, intrinsic moral value, moral accountability, privacy, and reciprocity. These contenders are attributed to a robot by the person interacting with it.

Coding of behaviors and deriving rules for interaction are another form of metrics adopted by some research groups. The problem when applying this procedure is the objective coding of behavior which actually is a subjective interpretation of an interaction scene as seen by an observer. In order to gain valid data the same interaction scenario should be coded by several independent observers of the experimental staff. So-called micro behaviors where used by [10] based on criteria like eye gaze, eye contact, operation and handling, movements, speech, attention, and repetitions. The length of eye gaze was used as a correlation to the subject's level of interest in a robot or toy truck. Behavior-level codes describing the adjustment of children to the setting of a communicative robot interacting with children in a primary school were used by [24, 21] to analyze the role of their robot.

So far many ideas, methodologies, metrics, and measurement criteria do exist in order to assess human-robot interaction, but most of the applied metrics consider mainly technical characteristics of the robot. Even success rates of interactions do not really picture the manifold ways of human behavior and the reasons for a failure of the interaction. The problem is that human behavior cannot be measured using simple scales. The assessment of interactions between naive persons and robots actually requires a framework of different metrics: a combination of objective metrics which can easily be measured and quantifiable subjective metrics characterizing human behavior. Here, undue influence of naive subjects as well as biased opinions of observers have to be taken into account by creating a set-up for sound experimentation and analysis.

In this paper we suggest contributions for a framework of metrics for human-robot interaction which considers the technical merits of the robot system as well as human behavior. Though this work does not present a full framework for evaluation of the whole robot, this work contributes aspects from different fields which have been applied to our robot.

The paper is organized as follows. Section 2 correlates robot skills in human environments with different levels of complexity, which demonstrates that different complexity levels also require different evaluation tasks. Section 3 describes example applications which provide a basis for the discussion of metrics, which is presented in Section 4. Section 5 concludes the article.

2. LEVELS OF COMPLEXITY

Relevant metrics for the evaluation of robot behavior crucially depend on the complexity of the task. In [8], we identified different levels of complexity for technical problems and cognition-related problems:

In Level 0, the robot only performs well-defined tasks defined by teaching or programming. The structure of tasks is known and the uncertainty about the environment is limited (e.g. positions of known objects). This level is state-of-the-art for most robots. The success of a task can be easily described by simple objective metrics like time duration or deviations from desired positions.

Level 1 tasks are characterized by a moderately higher degree of uncertainty. The robot can communicate with known human operators by understanding a fixed set of commands in natural language. The commands are associated with similar skills as outlined in Level 0. However, robots of this level need some slightly enhanced cognitive skills regarding error detection and handling. Most demonstrators for service robots are situated on this level. The necessary complexity of the metrics is also higher, e.g. for the description of the communication part.

Level 2 robots are designed for pre-defined domains in real-world environments like a kitchen. Besides many technical challenges for movement, communication and cognition aspects are extremely important, and the robot must be able to adapt to its environment. The interaction between robot and human being is mainly based on verbal and non-verbal communication. Here, the robot's active ability to converse and understand a naive person as well as its technical capabilities to overcome any acoustic diversions are of great importance. Close interaction with people also requires robots to respect a person's feeling of comfort. This includes the recognition of emotional states, the appropriate spatial distances between robot and person, the knowledge

about social cues when interacting with people etc. In addition, the robot should be able to learn to perform specific motions and actions by imitating a person demonstrating these movements. Consequently, all robot tasks depend on the recent scene and are difficult to evaluate.

Higher levels beyond the second one involve even higher complexity of interaction scenarios, environment, and required cognitive capabilities of the robot. First of all, the communication should not be restricted to a given domain. Different types of people and animals might come into contact with the robotic system. The robot should then know how to react to and interact with different types of people i.e. children, adults and handicapped people.

Such steps to higher levels are long-term goals whereas Level 2 robots should enable successful interactions between naive persons and robots in partially uncertain contexts.

From a metrics point of view, the evaluation of Level 2 and higher robots is much more sophisticated. In contrast to lower levels, the desired task is characterized by structural and parameter uncertainties. A possible manner is a parallel evaluation including a subjective and a objective component followed by a correlation of both parts. This solution will be outlined in the next section.

3. APPLICATIONS

3.1 Overview

This section describes different applications that address evaluation tasks from different disciplines and different aspects of a humanoid robot and its evaluation. The following aspects are addressed. Section 3.2 describes ideas for evaluation of human likeness concerning complex movements of the robot. Objective metrics are applied for quantification. In Section 3.3, two scenarios are discussed for evaluation of cognitive abilities in the area of interactive knowledge acquisition in dialog. Objective and subjective metrics are applied for quantification, and especially a trade-off is made between evaluation of single dialogs for knowledge acquisition and evaluation of the knowledge base. Section 3.4 presents a user study that has been conducted with video analysis and the use of qualitative measures. Requirements for experimental design are presented.

3.2 Quantification of Movements

Similarity measures for complex movements are a very important building block for complex metrics. They should cover the similarity of a robot movement to a desired human movement as well as the similarity of a human movement to a human reference movement. The first case is important for the evaluation of human-like robot behavior, the latter one for the on-line detection of human intentions by a robot.

In the last decade, computerized measurement systems for an instrumented motion analysis have become available [3]. Mostly, a camera-based tracking of external markers placed on a subject (Fig. 1) and a parameterized standard body model of a human being result in time series of joint angles. Newer systems try to overcome the need for external markers on subjects, but they are still under development. A robot only uses joint angles based on internal sensors as inputs.

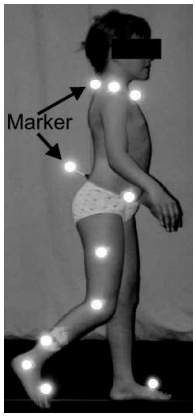


Figure 1: Child with markers for a motion analysis

A main application of such motion analysis systems is the clinical evaluation of movements e.g. for neurological diseases. Here, the calculation of quantitative measures is especially important for the monitoring of treatment progress [11, 2]. One option is the so-called reference distance of a set of L time series $x_{TS,l}[k]$ relative to a set of references. After a normalization of different time durations, it can be calculated by

$$x_{RD} = \frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{|x_{TS,l}[k] - \bar{x}_{Ref,l}[k]|}{\sigma_{Ref,l}[k]}, \qquad (1)$$

where $k = 1, \ldots, K$ are sample points of a time series with the length K, $x_{Ref,l}[k]$ is the l-th reference time series and $\sigma_{Ref,l}[k]$ is the corresponding standard deviation in the reference time series [34]. The standard deviation within the reference is used as a time-variant weighting factor. Consequently, a difference with respect to the reference is regarded as more severe when there is little variation within the reference motion pattern. A value of zero denotes an identity of the investigated time series to the reference, a value around one denotes a "normal" distance and larger values denote significant deviation. The main advantage of such a metric is the compression of a large amount of data to one scalar measure with a good interpretability. Nevertheless, a set (or different alternative sets) of reference time series is needed which can be a problem for complex movements.

As a clinical example, a group of 30 patients with an incomplete spinal cord injury was monitored over the course of their rehabilitation process resulting in a set of 81 analysis sessions [11]. They underwent treadmill training with partial body weight support as part of their therapy program. The reference data were obtained from a group consisting of 10 healthy subjects walking at comfortable self-selected speeds. Three independent clinical observers subjectively rated the quality of gait using video documentation on a Visual Analog Scale (VAS) from 0 (worst) to 10 (best). The correlation coefficients for the VAS values of the different observers were between 0.79 and 0.90. A reasonable correlation between VAS and objective metrics can be reached for the gait velocity (correlation coefficients to the observers: 0.68 - 0.79) and the reference deviation for the joint angles of ankle, knee, hip and pelvis in the sagittal plane measured by (1) (correlation coefficients to the observers: 0.60 - 0.67).

Such metrics allow an analysis of clinical observer's decisions and are a step to a more objective decision support system.

The alternative use of Hidden Markov Models (HMM) for similarity measures of motions based on a motion library [28] has advantages and disadvantages. On one hand, HMMs can model sequences of different motions including structural changes. On the other hand, they need a large set of training data and the interpretability of the internal processing is rather limited.

All discussed approaches offer the opportunity for the generation of objective evaluation metrics for the similarity between human and robot movements. Open problems include a sufficient representation of human reference movements for robots covering the variety of possible human-like movements during a robot task.

3.3 Knowledge Acquisition in dialog

Quite a number of dialog systems already exist and different metrics and evaluation schemes have been applied and tested. Metrics that have been applied to such systems are described in relevant literature as either objective or subjective metrics, or both in combination. As lined out before, both are useful for measuring. We first want to present objective metrics and then describe subjective user feedback.

3.3.1 Evaluation of dialog

The list of different objective metrics which have been applied to dialog systems is relatively short. Most systems use some kind of recognition accuracy, dialog length, and dialog success. Recognition accuracy can be represented as Word-Error-Rate (WER) which is the most simple metric. It has the advantage that it is usually used for evaluation and comparison of speech recognition systems and can easily be computed when the transcription of speech input is given. However, WER is not necessarily the best metric to represent recognition accuracy. For example, it doesn't distinguish between content words and non-content words. Sentence-Error-Rate (SER) checks the correctness of complete sentences. Some evaluations measure correctly recognized semantic concepts, for example (semantic) Concept-Error-Rate (CER) [9, 12, 15]. Differences exist whether CER is defined on fully correct semantic input or regarding the details used to measure correctness. CER is probably the metric, which is best suited to represent input understanding in a dialog system, because it is measured by the correctness of the input, which is actually used by the dialog manager. However, it requires semantic transcription of input, and is not as simple as word-error rate, since it depends on the type of semantic structure and details of semantic transcription.

dialog length is usually measured in number of turns to achieve a certain goal. In task-oriented systems the number of turns is measured to achieve a predefined task. Some other metrics have been used, such as the total amount of time in seconds, or the number of syllables spoken [27]. [12] uses concept efficiency (CE) which quantifies the average number of turns necessary for each concept to be understood by the system, and query density (QD) which measures the mean number of new concepts introduced per user query. Both metrics relate to the length of the dialog with respect to how effectively information can be communicated without the necessity of a task definition.

A widely used metric is dialog success. However, the definition of dialog success varies among different systems. Most approaches use achievement rates of dialog goals, e.g. [26].

As a framework for dialog system evaluation, PARADISE [32, 31] is best known. It offers a prediction model for quality judgments based on a regression model with interaction parameters as input. It serves two purposes, one part is the framework for prediction of quality judgments, the second part is a set of questions and metrics for evaluation. The framework has been applied to a number of different systems, for example [13]. Since PARADISE has initially been designed for speech-only interactions, a modified version, PROMISE, has been suggested by [4] to address aspects of multimodal systems.

Such frameworks apply both, objective and subjective metrics. Subjective evaluation is usually conducted with the help of questionnaires, which allow quantitative measurements based on Likert-Scales. A Likert-scale is a unidimensional scale with a discrete set of response possibilities, usually a 5-point, or sometimes a 7-point scale to rate between Disagree and Agree. Questions are then formulated as statements. Some approaches use different opposites than agreement or disagreement, such as *'good'* vs. *'bad'*, *'very much'* vs. *'not at all'*. Questions are then formulated as real questions, such as "How is your overall impression of the interaction?". An analysis of de-facto evaluation standards for quality of the interaction with spoken dialog systems is presented in [23]. Here, two questionnaire methods are compared, the SASSI questionnaire [18, 17] and the ITU-T Recommendation P.851. In addition, a classification scheme and taxonomy of quality aspects were presented. As a result, both questionnaires provide valid measurements of different quality aspects. The subjective feedback was combined with extracted parameters to predict system usability and acceptability. Extracted parameters are integrated into a prediction model with the PARADISE framework, from which helpful information was obtained for system design, but not general predictions of system usability and acceptability.

3.3.2 Application to Object Learning

Advanced humanoid robots, as we have described in the previous sections as Level 2 and Level 3 robots, need to be able to adapt to their environment, which includes that they are able to learn new information. For learning we describe two tasks. The first task addresses learning of objects, the second task addresses interactive learning of person ID models.

Interactive learning is a key requirement for cognitive autonomous systems. A humanoid robot, e.g. in a household environment frequently encounters previously unknown objects. We have addressed this task with a scenario, where the robot can interactively learn new objects, such as CDs, books, chocolate, etc. Our approach for interactive learning of objects integrates several knowledge sources and aspects:

- *visual information* is acquired and stored for new objects for visual recognition

task	#dialogs	success	avg turns
learn object property	40	83% (33)	1.8
- with known words	25	87% (22)	1.4
- with spelling	15	74% (11)	2.6

Table 1: Success rates and dialog length for learning of object properties. Numbers are given for evaluation on all dialogs, and separated by whether words were known to the robot, or spelling has been used.

- *different descriptions for reference in speech* can be acquired for a new object, which covers introduction of new words
- *semantic information* about the object is acquired in dialog. Semantic information covers the type of the object and properties.

Quantification of successful interactions and success of knowledge acquisition can be defined on a dialog basis (evaluation of the interaction) or by means of knowledge base quality. Metrics for evaluation of the knowledge base can be defined in terms of correctly learned objects on a predefined set of objects which are shown to the robot. Additionally, we suggest to use recognition rate of learned objects, to evaluate if the learned knowledge is applicable.

These metrics were applied in a small system evaluation with 52 dialogs, during which some of these metrics were studied, regarding system components, dialog-based metrics, and quality of the knowledge base. 40 dialogs were conducted with unknown objects. Component evaluation measures recognition rates of object recognition and speech recognition. dialog-based metrics consider dialog length and dialog success. In this evaluation, dialog success also represents the learning rate. 12 dialogs were conducted with previously learned objects. Here, quality of the knowledge base was evaluated in terms of recognition accuracy for learned objects, which is only one possibility described above to assess quality of the knowledge base. During all dialogs, recognition rates were calculated for all attempts to recognize an object, which is the standard evaluation metric for recognition components, and recognition rates including confirmation dialogs. Standard recognition rate e.g. is 81% for visual recognition. Including confirmation dialogs and a second attempt to recognize the objects, the recognition rate is 94%. Success of the learning dialogs is measured in success rates to obtain object descriptions, object type and object properties. Table 1 shows numbers for learning of dialog properties from the experiment. In a similar way, the dialogs for learning of semantic categories have been evaluated. For comparison of different dialog strategies, different tests are conducted were the strategies are applied iteratively.

3.3.3 Application to Person ID Learning

Our scenario for interactive learning of person ID information is based on an interactive system in a corridor as a robot receptionist. Its task is to greet persons who pass the robot, engage in a dialog, try to identify the person and once the person is known obtain more personal information. First studies and experiments for identification have already been conducted [14, 16], which evaluate success of engagement and dialogs for learning names, and optimization of dialog strategies with reinforcement learning.

Evaluation of this scenario again integrates evaluation of the dialogs and evaluation of the knowledge base. A series of experiments has already been conducted with this scenario. First evaluations have also been conducted, but a more profound evaluation with all metrics suggested here, is still to be realized. For evaluation of the knowledge base we first consider the ground truth of pieces of information that can be learned by the system. Ground truth must be known in advance or be transcribed from collected data. Information gathered by the system and which constitutes the knowledge base are the ID of a person, the person's first and last name, a portrait snapshot of the person and several attributes associated with the person. Attributes associated with the person are pieces of information like the email address, social relations with other persons, and research topics. Each of these items can be evaluated with binary values *'exists'* vs. *'doesn't exist'*, *'correct'* vs. *'wrong'* - except social relations for which other metrics need to be defined. Based on these single binary values we plan to evaluate the knowledge base with standard metrics used in machine learning, from which especially precision/recall provides a good metric.

In [7] we have presented a user study that was conducted as a Wizard-of-Oz experiment. The system was designed as a standalone system, and a human operator took over the dialog decisions. The system was tested with 16 persons, each of them interacting with the system on three consecutive days. Here, differences between dialog success and knowledge base quality become obvious. To measure dialog success, the number of dialogs is counted, in which the person was correctly identified. It must be noted that the success metric can be applied in different ways. For example, some names are phonetically identical such as Stephan and Stefan, thus there are differences to which degree information must be correct to account for a successful dialog. The numbers of successful dialogs, where success defines correct identification and acceptable name pronunciation, was 56% at day 1, 81% at day 2, and 100% at day 3. On average these are 79% successful dialogs. The knowledge base quality is slightly different. Using the same criteria as before, recall is 56% by the end of day 1, 88% by the end of day 2, and 100% by the end of the experiment. In addition, precision is calculated, which is 73% by the end of day 3, since 3 persons have been stored incorrectly and 3 other persons have been stored twice, with similar names (e.g. as Chilipp and as Philipp).

3.4 Dialog-based Human Robot Interaction

In order to assess and compare various forms of human-robot-interaction we conducted several experiments featuring rather simple scenarios like a bartender robot, a receptionist robot with person ID learning and a communicative robot with person ID learning and knowledge acquisition about social networks (Fig. 2). In all experiments the subjects had to talk to the robot, either to receive or to give information.

A multitude of quantitative and qualitative methods of empirical social research were used for the sociological evalua-

Figure 2: Person interacting with the robot.

tion of all experiments. Protocols were written by a team member during each interaction, invisible for the subjects; after each experiment questionnaires had to be filled out and interviews were conducted. Quantitative data were gained by the questionnaires and the protocols: the first supplied the answers of open questions based on scales of seven increments, the latter supplied data about the number of turns, total duration or number of breakages in connectivity. The bartender robot as well as the network-building robot both only talked in English whereas the receptionist robot talked in German. Some difficulties resulted from English not being the native tongue of the subjects.

The evaluation itself focused activities of the actors as well as the interaction, which were recorded on videos from two different angles (an overall view of the scene as well as a frontal recording of the face and upper body of the subjects). A special tool called Interaction Analysis Tool (IAT, formerly called IAP [7]) has been developed by our team for the quantitative as well as the qualitative analysis of the recorded interactions. There are several layers; the first pictures what happened, when and where during an interaction. This is a typical transcript of video data, which can be conducted by any social scientist skilled in objective transcription of videos, thus supplying data for the quantitative analysis. The other layers provide qualitative data, as here phases of the interaction, specific events, and criteria (i.e. initiative, coherence, transparency, redundancy, information strategy) are emphasized (for a close description of the IAT (IAP) please refer to [7]). Why and how something happened is addressed by these layers; this is an interpretation and hermeneutic description of the observed actions (behaviors), which is best done by a group of evaluators independently. In this way the meaning of a sequence of actions or turns can be better explained by following the best arguments of the independent evaluators. Additional qualitative data are gained by extracting the subjects' stories about the robot and the situation from the interviews. These stories often corroborate an interpretation of a sequence of actions. The actual criteria of an interaction, as defined within the IAT, as well as specific events (loops, omissions, breaks, ...) and, phases, form patterns which can be linked to a specific behavior of either robot or subject. In this way the human factor can be assessed far more explicitly than just considering the mere success rate of an interaction.

4. DISCUSSION

The previous sections have shown applications of different objective and subjective metrics for different tasks. In this section, we propose first ideas of a generic concept for the evaluation of Level 2 robots. It includes remarks to

- the experimental design,
- the integration of existing objective metrics, and
- steps to a quantification of subjective metrics.

To obtain reliable results, sound **experimental design** is essential in order to avoid systematic errors. Here, insights from sociology and psychology need to be considered. First, appropriate test scenarios are necessary. To cope with the uncertainties resulting from the complex environment, the tests scenarios should be designed as a series of tests on different test trials.

For our experiments with dialog based human robot interaction, we have selected persons that were completely unknown for the robot at the first meeting. In addition, the subjects have not been prepared about the robot's capabilities before interaction with the robot. While conducting experiments with people who are familiar with human-robot interaction can solely evaluate interactions of persons that are familiar with the robot, experiments with 'naive' users are the only way to model situations in which people first meet a robot. It is a natural situation for the first "getting in contact" with personal robots or meeting unknown guests during the whole lifetime of a robot. Generally, such meetings should be performed for a higher number of different subjects (e.g. age, sex, profession) to catch the bandwidth of human communication behaviors. Such tests evaluate the robustness of the robot to communicate with different subjects characterized by completely different expectations about the robot's behavior. The environment of the experiment needs to be taken special care of, with the target scenario in mind. For example to represent an everyday scenario in a household environment would contain background noise e.g. with radio, TV, or additional people in the background. The robot operators should not be visible to avoid an influence by tips and hints to the interacting persons.

In addition, repetition trials with the same persons at different days can imitate an alternating learning process between human and robot. Here, the robot's ability to exploit the knowledge about a special subject to adapt its own behavior is tested.

Test subjects have different notions about the success or failure of an interaction as well as the reasons for failure and build up different internal mental pictures of the robot's capabilities. To obtain these notions, subjective user feedback can be obtained in form of a user questionnaire to obtain quantifiable results or in form of an interview to obtain qualitative feedback.

Objective metrics can mostly be calculated automatically and need no further processing steps in between by a person.

They are very useful in many parts of the system. However, with the current knowledge about metrics for human-robot interaction they are not sufficient. Quantitative, objective metrics evaluate e.g. success rates for underlying skills like navigation (e.g. possible speed, deviation from the planned route, number of touched obstacles), speech recognition and generation (e.g. number of successful dialogs), manipulation (e.g. number of successful handled objects). They aggregate metrics from underlying low-level motor and perceptions skills with related metrics such as tracking errors to planned joint trajectories, control errors, and stability margins of basic control loops, maximal movement speeds for robot joints, the number of not-identified objects, localization accuracy of objects according distance, height, number of recognized words, delay time for answers etc. Some of these metrics are useful for self-evaluation of the robot to improve its abilities for adaptation, fault detection, and safety-related supervision.

A guiding principle to design quantitative on-board metrics is the evaluation of differences between the expected and actual behavior of robot and environment. These metrics are good measures for the ability of the robot to predict future situations based on its own experiences.

All metrics listed above are useful to evaluate some facets of the robot behavior. It will be impossible to aggregate all metrics to an overall performance metric due to their different natures, their requirements in further processing and interpretation and their partially non-automatic acquisition. An alternative will be a multi-modal evaluation leading to scores for Pareto-optimal robot behaviors with advantages for different sub-metrics.

Subjective metrics like questionnaires, interviews, stories, protocols taken by team members, and video data are hard to quantify and qualify. They vary according to the people involved. Usually, there are no measurable scales on comfort, acceptance, or other criteria, as each test subject has his or her own scale as well as his or her individual image of the interacting robot. Additionally, video data, questionnaires, and log-files have to be compared, as statements given in questionnaires can contradict the actual behavior of a test subject recorded on video.

Also the coding or evaluator's side has different amount of subjective influence, depending on the evaluation task. This is not wanted to obtain objective measurements, and thus needs to be taken special care of. As soon as additional people are involved in coding human behavior, whether actions, reactions or adjustments from video data, different personal views enter the coding. Actually, the same video sequence should be coded by at least two different people.

All subjective metrics need additional people to either process the data i.e. transcription of video files or coding or to analyze and interpret acquired data. Here again, personal opinions of the evaluators can impair results. Once either category has been assigned to individual test runs and turns between a person and a robot or behavior has been coded, these codes can serve as metrics, as coded categories and behaviors form characteristic patterns [6, 10, 24]. Then different test runs with the same subject as well as test runs with different subjects are much easier to assess and compare.

In the literature of recent work in evaluation of human-robot interaction we have found the following approaches for subjective metrics:

- questionnaires, the outcome depends on the kind of questions and scales
- interviews, the outcome depends on the kind of questions and the interviewer,
- video analysis, which requires video transcription. Due to differences in coding, several people should code the same sequences,
- coding of behavior [24], [10],

[6] describes the following criteria for objective assessment based on coded subjective criteria

- criteria describing the interaction context (interaction patterns, interaction rules, roles, degree of freedom),
- criteria describing the interaction itself (intensity, congruity, convergence, synergy, efficiency),
- criteria describing the activity of actors (transparency, roles),
- criteria describing non-verbal actions and emotions (mimics, postures, gestures, affects),
- specific coding of micro behaviors.

Another idea is a linear or nonlinear correlation of existing subjective and objective metrics for a larger set of experiments. This concept might contribute to a deeper understanding of human impressions coded by the subjective metrics. If strong correlations between such metrics can be found, the related objective metrics might be used as independent variables in a regression model for the description of subjective metrics as dependent variables. For this approach, a careful statistical analysis is necessary to avoid an over-fitting of a model based on a small number of experiments.

5. CONCLUSION

This paper analyzes metrics for human-like movements and interaction patterns for intelligent robots. Such metrics are needed for the evaluation of humanoid robots interacting with people in a natural environment. Existing quantitative metrics for robot evaluation are often focused on a small subset of technical robot skills. This kind of evaluation does not cover important aspects for complex scenarios like the overall performance in a task or the comfort and acceptance for interacting people. On the other hand, many proposed subjective metrics like questionnaires do not result in quantitative scales needed for an objective evaluation of human-robot interaction.

We propose the integration of strategies for experimental design and evaluation criteria coming from a technical, psychological and sociological background. Different aspects are discussed in the context of applications, including metrics for the similarity of human movements, knowledge acquisition in dialog and the dialog-based human robot interaction. The experimental design must be oriented on the robot's expected future field of application. If e.g. the robot will mainly interact with unknown persons, the test scenario should base on the interaction with a larger group of naive subjects without specific knowledge about the robot capabilities. Only such test scenarios will provide realistic evaluations of robot behavior. Information from qualitative metrics should be step-wise quantified, e.g. by means of correlation analysis between questionnaires and candidates for quantitative features.

A main problem of quantification in HRI analysis is the transcription and coding of recorded data from experimental runs as a sequence of human and robot actions. Up to now, it requires a huge manual effort. These results might be valuable building blocks for quantitative metrics. An automation of these steps, i.e. complete reliance on objective metrics, is only possible if the robot is able to identify and to interpret human behavior by itself. Till then, a mixture of objective and subjective metrics, quantitative and qualitative evaluation constitutes a necessary mixture of different aspects for evaluating robot performance in human-robot interaction scenarios.

6. ACKNOWLEDGMENTS

This work has been performed within the framework of the German humanoid robotics program (SFB 588) funded by the German Research Foundation (*DFG: Deutsche Forschungsgemeinschaft*).

7. REFERENCES

[1] DARPA urban challenge - event guidelines. Defense Advanced Research Projects Agency (DARPA), Arlington, 2007.

[2] R. Abel, R. Rupp, and D. Sutherland. Quantifying the variability of a complex motor task specifically studying the gait of dyskinetic CP children. *Gait & Posture*, 17:50–58, 2003.

[3] R. Baker. Gait analysis methods in rehabilitation. *Journal of Neuroengineering and Rehabilitation*, 3:4:1–10, 2006.

[4] N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Türk. Promise - a procedure for multimodal interactive system evaluation. In *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation'. Las Palmas, Gran Canaria, Spain*, 2002.

[5] A. Bredenfeld, A. Jacoff, I. Noda, and Y. Takahashi, editors. *RoboCup 2005: Robot Soccer World Cup IX*. Springer, Lecture Notes in Computer Science, 2006.

[6] C. Burghart and R. Haeussling. Evaluation criteria for human robot interaction. In *Workshop on Robotic Companions, AISB 2005*, Hatfield, England, Apr. 2005.

[7] C. Burghart, H. Holzapfel, R. Haeussling, and S. Breuer. Coding interaction patterns between human and receptionist robot. In *Proceedings of Humanoids*, 2007.

[8] C. Burghart, R. Mikut, and H. Holzapfel. Cognition-oriented building blocks of future benchmark scenarios for humanoid home robots. In *Proc., AAAI Workshop Evaluating Architectures for Intelligence, Vancouver*. 2007.

[9] A. Chotimongkol and A. Rudnicky. N-best speech hypotheses reordering using linear regression. In *Proceedings of Eurospeech*, 2001.

[10] K. Dautenhahn and I. Werry. A quantitative technique for analysing robot-human interactions. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1132 – 1138. 2002.

[11] J. Dieterle. *Bewertung der Gangqualität inkomplett querschnittgelähmter Patienten mit Instrumentellen Ganganalysen*. PhD thesis, Medizinische Fakultät Heidelberg der Ruprecht-Karls-Universität, 2006.

[12] J. Glass, J. Polifroni, S. Seneff, and V. Zue. Data collection and performance evaluation of spoken dialogue systems: The mit experience. In *Proc. of ICSLP*, Beijing, China, October 2000.

[13] M. Hajdinjak and F. Mihelič. The paradise evaluation framework: Issues and findings. *Computational Linguistics*, 32(2):263–272, June 2006.

[14] H. Holzapfel, T. Schaaf, H. K. Ekenel, C. Schaa, and A. Waibel. A robot learns to know people - first contacts of a robot. *Lecture Notes in Computer Science - KI 2006: Advances in Artificial Intelligence*, 4314, 2007. Freksa, C., Kohlhase, M., Schill, K. (eds.).

[15] H. Holzapfel and A. Waibel. A multilingual expectations model for contextual utterances in mixed-initiative spoken dialogue. In *Interspeech 2006 - ICSLP*, Pittsburgh PA, USA, 2006.

[16] H. Holzapfel and A. Waibel. Behavior models for learning and receptionist dialogs. In *Interspeech 2007*, Antwerp, Belgium, 2007.

[17] K. Hone and R. Graham. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3/4):287–305, 2000.

[18] K. Hone and R. Graham. Subjective assessment of speech-system interface usability. In *Proceedings of Eurospeech*, pages 2083–2086, 2001.

[19] A. Jacoff, E. Missina, and J. Evans. Performance evaluation of autonomous mobile robots. *Industrial Robot: An International Journal*, 29(3):259–267, 2002.

[20] P. Kahn, H. Ishiguro, B. Friedman, and T. Kanda. What is a human? toward psychological benchmarks in the field of human–robot interaction. In *Proc., IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 2006.

[21] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1-2):61–84, 2004.

[22] A. Lampe and R. Chatila. Performance measure for the evaluation of mobile robot autonomy. In *Proc, IEEE International Conference on Robotics and Automation (ICRA'06)*, pages 4057–4062. 2006.

[23] S. Möller, P. Smeele, H. Boland, and J. Krebbera.

Evaluating spoken dialogue systems according to de-facto standards - a case study. *Computer Speech & Language*, Volume 21, Issue 1:26–53, January 2007.

[24] S. Nabe, S. J. Cowley, T. Kanda, K. Hiraki, H. Ishiguro, and N. Hagita. Robots as social mediators: coding for engineering. In *Proc. of the International Symposium on Robot and Human Interactive Communication (Ro-Man)*, Hatfield, UK, 2006.

[25] D. Nardi and et al. *RoboCup@Home: Rules and Regulations (Draft, Version 1.0, Revision 16).* RoboCup Federation, 2007.

[26] J. Schatzmann, K. Georgila, and S. Young. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal, 2005.

[27] G. Skantze. Making grounding decisions: Data-driven estimation of dialogue costs and confidence thresholds. In *Proceedings of SigDial*, pages 206–210, Antwerp, Belgium, 2007.

[28] T. Stein, A. Fischer, K. Bös, V. Wank, I. Boesnach, and J. Moldenhauer. Guidelines for motion control of humanoid robots: Analysis and modelling of human movements. *International Journal of Computer Science in Sports*, 5:15–30, 2006.

[29] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. Common metrics for human-robot interaction. In *Proc., ACM SIGCHI/SIGART Human-Robot Interaction*, pages 33–40. ACM Press New York, NY, USA, 2006.

[30] S. Thrun and et al. Stanley: The robot that won the DARPA grand challenge. *Journal of Field Robotics*, 23(9):661–692, 2006.

[31] M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with paradise. *Nat. Lang. Eng.*, 6(3-4):363–377, 2000.

[32] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Paradise: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. ACL*, pages 271–280, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[33] M. L. Walters, K. Dautenhahn, S. N. Woods, and K. L. Koay. Robotic etiquette: Results from user studies involving a fetch and carry task. In *Proc. of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 317–324, New York, NY, USA, 2007. ACM.

[34] S. Wolf, T. Loose, M. Schablowski, L. Döderlein, R. Rupp, H. J. Gerner, G. Bretthauer, and R. Mikut. Automated feature assessment in instrumented gait analysis. *Gait & Posture*, 23(3):331–338, 2006.

[35] S. Woods, M. Walters, K. Koay, and K. Dautenhahn. Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach. In *Proc. of the International Workshop on Advanced Motion Control*, Istanbul, Turkey, 2006.

Identifying Generalizable Metric Classes to Evaluate Human-Robot Teams

P. Pina
Massachusetts Institute of Technology
Cambridge, MA

M. L. Cummings
Massachusetts Institute of Technology
Cambridge, MA

J. W.Crandall
Massachusetts Institute of Technology
Cambridge, MA

M. Della Penna
Delft University of Technology
The Netherlands

ABSTRACT
In this paper, we describe an effort to identify generalizable metric classes to evaluate human-robot teams. We describe conceptual models for supervisory control of a single and multiple robots. Based on these models, we identify and discuss the main metric classes that must be taken into consideration to understand team performance. Finally, we discuss a case study of a search and rescue mission to illustrate the use of these metric classes to understand the different contributions of team performance

Categories and Subject Descriptors
J.7 [**Computers in Other Systems**]: Command and Control; H.5.2 [**User Interfaces and Presentation**]: Evaluation/ methodology

General Terms
Measurement, Performance, Experimentation, Human Factors

Keywords
Metrics, Human-Robot Teams, Performance, Supervisory Control

1. INTRODUCTION
Mission effectiveness is the most popular metric to evaluate the performance of human-robot teams. However, frequently this metric is not sufficient to understand team performance issues and to identify design improvements, and additional metrics are required.

Despite the importance of selecting the right metrics, few general guidelines that apply to a wide range of human-robot applications are available in the literature. In many cases, researchers rely on their own experience, selecting metrics they have used previously. Alternatively, other experiments measure every system parameter to ensure that every aspect of system performance is covered. These approaches lead to ineffective metrics and excessive experimental and analysis costs. Moreover, existing metrics for evaluating human-robot teams are usually application-specific, which makes comparison across applications difficult.

The goal of this research is to provide general guidelines for metric selection that are applicable to any human-robot team operating under a supervisory control paradigm. We believe that identifying generic metric classes that organize the different types of metrics available will help researchers select a robust set of metrics that provide the most value for their experiments and allow comparison with others. Metrics may still be mission-specific, however metric classes are generalizable across different missions. In the context of this paper, a metric class is defined as the set of metrics that quantify a certain aspect or component of a system.

The idea of developing a toolkit of metrics and identifying classes to facilitate comparison of research results has already been discussed by other authors. For example, Olsen and Goodrich proposed four metric classes to measure the effectiveness of robots: task efficiency, neglect tolerance, robot attention demand, and interaction effort [1]. This set of metrics measures the individual performance of a robot, however, a particular robot performance does not necessarily imply a level of human performance. Since human cognitive limitations often constitute a primary bottleneck for human-robot team performance, a metric framework that can be generalized should also include cognitive metrics to understand what drives human behavior and cognition.

In line with this idea of integrating human and robot performance metrics, Steinfeld et al. suggested identifying common metrics for human-robot interaction in terms of three aspects: human, robot, and the system [2]. Regarding human performance, they discussed three main metric categories: situation awareness, workload, and accuracy of mental models of device operations. This work constitutes an initial step towards developing a metric toolkit, however it still presents some limitations. On the one hand, this framework suffers from a lack of metrics to evaluate collaboration effectiveness among humans and among robots. On the other hand, a more comprehensive discussion on human performance is still required. For example, the authors discuss trust as a task-specific metric for social robots but it is not included as a common metric required to evaluate operator performance. We believe that operators' trust in robot behavior is often a key factor in team performance.

The research presented in this paper builds upon previous efforts conducted by Crandall and Cummings [3]. It refines, expands, and generalizes the set of metric classes already identified for human-robot teams consisting of a single human and multiple robots. The paper builds a conceptual model for human supervisory control of multiple robots. Then metric classes are identified from this model. Finally, a case study on a search and rescue mission is discussed to illustrate some of the proposed metric classes.

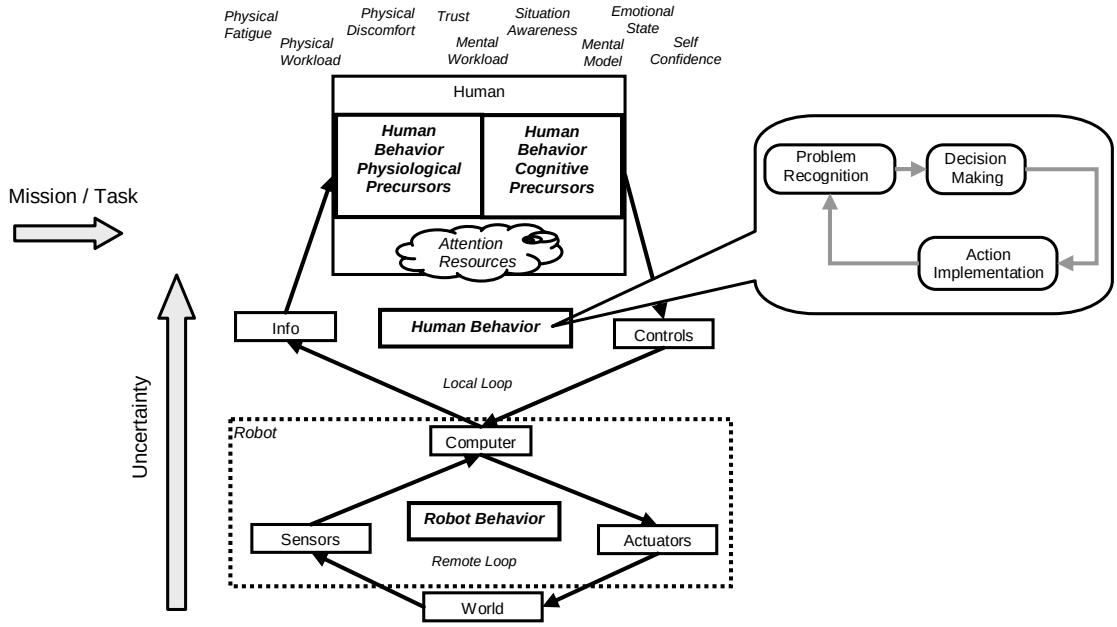


Figure 1. Conceptual Model of Human-Robot Interaction in Supervisory Control.

2. CONCEPTUAL MODEL

This section presents and discusses our conceptual models of human supervisory control of robots, including a single operator controlling a single robot, a single operator controlling multiple robots, and multiple operators controlling multiple robots.

2.1 Supervisory Control of a Single Robot

"Supervisory control means that one or more human operators are intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors and sensors to the controlled process or task environment [4]." Most human-robot teams operate under a human supervisory control paradigm where robots have a certain degree of autonomy and the human guides them, monitors their performance, and intervenes when needed. Examples of this are found across several domains and applications: surveillance and target identification for military operations, health care applications such as mobility assistance and therapy, rock sampling for geology research, or other logistic applications such as personnel or material delivery.

All these examples can be conceptually represented by the model shown in Fig. 1. This model is composed of four interrelated main elements: robot behavior, human behavior, human behavior cognitive precursors, and human behavior physiological precursors. We believe that these four elements delineate the main metric classes for single operator-single robot teams. In addition to these four elements, two other concepts are represented in Fig. 1: uncertainty, and the mission or the task. Uncertainty refers to the uncertainty associated with sensors (e.g., accuracy) and actuators (e.g., lag), displays (e.g., transforming 3D information into 2D information), and the real world. This uncertainty propagates through the system reaching one or more operators who adapt their behavior to the uncertainty level by applying different cognitive strategies.

Regarding the mission or the task imposed on the operator, human behavior and system performance depend on the nature of the tasks. High structured tasks, those that can be planned in advanced and are procedurally-driven, are very different, from a human perspective, from those that have low structure levels, which are generally emergent tasks that require solving a new problem under time-pressure. Human-robot team performance can only be understood if considered in the context of the mission and the task.

The goal of this paper is to develop a general framework for the analysis of human-robot team performance. However, our focus is on those metrics of human behavior efficiency and human behavior precursors, rather than metrics of robot behavior efficiency. The fact that many human-robot teams are remote makes it essential to measure the human component. Operators who remotely operate a robot do not physically perceive the interaction of the robot with the real world. This can have a negative impact on situation awareness and human trust, which in turn can affect performance.

2.1.1 Robot & Human Behavior Efficiency

Robot and human behavior are represented by the two control loops shown in Fig.1: the human control loop and the robot control loop. The operator receives feedback on robot and mission performance, and adjusts robot behavior through controls if required. The robot interacts with the real world through actuators and collects feedback on mission performance through sensors. The evaluation of team performance requires an understanding of both control loops. The rest of this section focuses on human behavior.

Human behavior, in the context of Fig. 1, refers to the decisions made and actions taken by the human while controlling the robot. The model presented in Fig. 1 categorizes human behavior in terms of problem recognition, decision making, and action implementation. These three categories are based on the four-stage model of human information processing described by Parasuraman, Sheridan, and Wickens: 1) information acquisition, 2) information analysis, 3) decision and action selection, and 4) action implementation [5]. Our model merges the stages of information acquisition and analysis into the problem recognition category. Acquisition and analysis of information are often hard to differentiate, and the human

ability to recognize problems is a more valuable metric for our purposes. Thus, understanding human performance requires evaluating each one of the three categories defined by our model.

Human-computer interactions (HCIs) are the observable outputs of human decisions, and they are commonly used to measure human behavior efficiency. Based on our model, these interactions should also be analyzed in terms of problem recognition (e.g., access to information about the environment dynamics), decision making (e.g., use of what-if functionalities to explore consequences of actions), and action implementation (e.g., entering new coordinates for a robot's destination). Such decomposition enables a more comprehensive evaluation of team performance. However, disaggregating HCIs may not always be possible.

In addition to human efficiency for problem recognition, decision making, and action implementation, human attention allocation is a key component of human behavior. The evaluation of attention resource allocation helps in the understanding of operators' strategies and priorities. Operators have limited attention resources that need to be shared between multiple tasks [6]. Although as seen in Fig. 1, one single robot is controlled, the operator still performs multiple tasks such as monitoring the dynamics of the environment, identifying emergent events, monitoring robot health, or executing manual control of the robot. How humans sequence and prioritize these multiple tasks provides valuable insights into the system.

2.1.2 Human Behavior Cognitive and Physiological Precursors

Evaluating human observable behavior can still be insufficient since all mental processes do not have immediate and observable outcomes. The evaluation of human performance requires understanding what motivates the behavior and the cognitive processes behind it. Human behavior is driven by high level cognitive constructs and processes such as mental models[1] and situation awareness[2] (SA). For our discussion, mental models refer to long-term knowledge, whereas SA reflects dynamic knowledge. Understanding human mental models is important because ideally, an interface design should be consistent with people's natural mental models about computers and the environment [10]. Poor SA or lack of understanding of a dynamic environment, when performing complex cognitive tasks, can have dramatic consequences such as the incident at Three Mile Island [11].

Mental models and SA are not the only human behavior cognitive precursors. In the context of this paper, human behavior cognitive precursors refer to cognitive constructs or processes that existed or occurred before a certain behavioral action was observed. Human trust in the robots, mental workload, and operator emotional state are other examples of cognitive constructs and processes that can also cause certain human behaviors.

Furthermore, physiological processes can reflect physical states such as fatigue, or physical discomfort which can also motivate certain human attitudes.

2.1.3 Conclusions

Our model represents the need for evaluating four main elements to understand the performance of a single operator-single robot team: robot behavior, human behavior, human behavior cognitive precursors, and human behavior physiological precursors. These four elements are all interrelated. For example, events in the real world are captured by the robot sensors and presented to the human operator through the display. Modifications on the display can affect human attention allocation and SA, which in turn will result in changes in HCI patterns, which can ultimately affect robot performance. Understanding system performance implies understanding the relations among these elements.

2.2 Supervisory Control of Multiple Independent or Collaborative Robots

The previous section discusses a model for one operator-one robot team, but operators can simultaneously control multiple robots. In order to expand the model in Fig. 1, we consider two different scenarios: a) multiple robots performing independent tasks, and b) multiple robots performing collaborative tasks. In this paper collaboration between robots means two or more robots working together to accomplish a shared goal under human supervision.

In the case of independent robots, servicing robot 1 and robot 2 are two independent tasks. The operator monitors the environment and the robots, decides on which one to focus his/her attention, interacts with that robot, and returns to monitoring or decides to service another robot. While servicing one of the robots, the operator behaves similarly as if he/she supervised only one single robot. Our model assumes that the operator does not service multiple robots in parallel. This assumption is based on the limited human cognitive resources and the high task demands imposed by supervising complex and dynamic environments under time pressure. Figure 2 illustrates this model of human supervisory control of multiple independent robots.

[1] The phrase "mental models" refers to organized sets of knowledge about the system operated and the environment that are acquired with experience [7].

[2] SA is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [8]. In the context of human-robot teams, SA encompasses awareness of where each robot and team member is located and what they are all doing at each moment, plus all the environmental factors that affect operations [9].

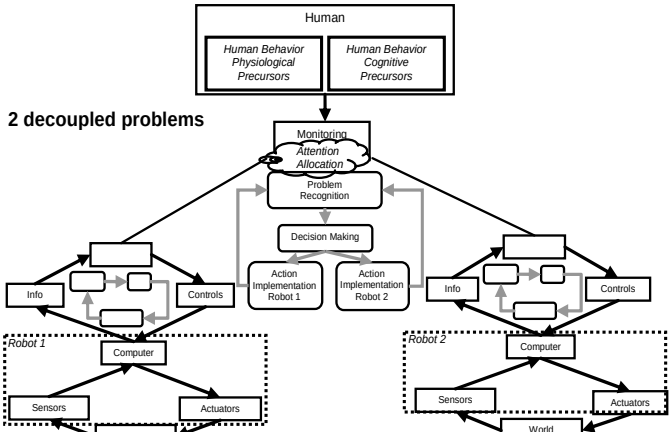


Figure 2. Supervisory Control of Independent Robots.

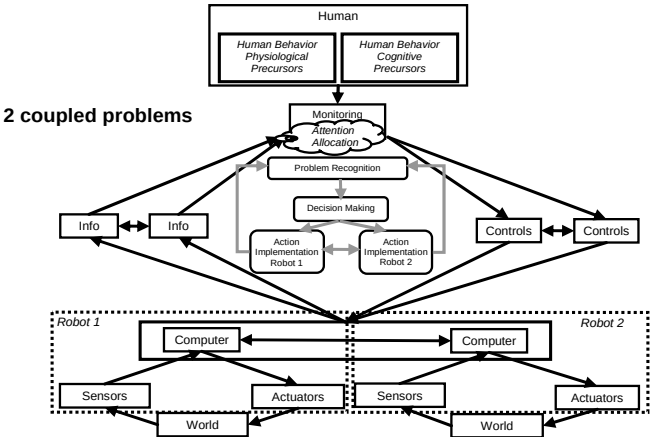


Figure 3. Supervisory Control of Collaborative Robots.

Multiple robots working together to achieve a common goal can autonomously collaborate or be manually coordinated by the operator. In the case of autonomous collaboration among robots without the possibility for human intervention, collaboration only occurs at the level of the robot behavior loop and the model in Fig. 2 is still valid. However, in the case of active human coordination, the operator executes two dependent tasks (i.e., servicing robots 1 and 2) that cannot be decoupled. Figure 3 illustrates the later model, where the control loops for robot 1 and robot 2 are not independent and separated entities. Controlling collaborative robots requires the operator to understand the consequences of an action across both control loops and to actively coordinate between them. For example, making a decision for robot 1 can involve acquiring and analyzing information related to robot 2, and implementing an action for robot 2 can require synchronizing it with another action for robot 1. Interfaces for collaborative robots should aggregate data from each control loop and display it so that the operator can easily understand the interconnections and the consequences of these dependencies.

In our previous example with independent robots, the three categories of human behavior (i.e., problem recognition, decision making, and action implementation) could be evaluated separately for robot 1 and robot 2. In the case of collaborative robots, these three categories have to be analyzed for both robots aggregately.

2.3 Human Collaboration in Supervisory Control of Multiple Robots

This section expands previous models to the case of multiple humans collaborating to control multiple robots. In these situations, system performance is directly linked to human collaboration. Our model considers two main dimensions of collaboration: team behavioral actions and team cognition. Figure 4 illustrates this model.

The evaluation of team behavioral actions consists of measuring both the efficiency of team coordination and the team efficiency in each of the three categories of human behavior (i.e., problem recognition, decision making, and action implementation). The team works together as a single entity to perform collaborative tasks so performance should be measured at the holistic level rather than aggregating team members' individual performance [12]. Team coordination comprises of written, oral, and gestural interactions among team members.

Team cognition refers to the thoughts and knowledge of the team. Measures of team cognition can be valuable in diagnosing team performance successes and failures, and identifying training and design interventions [12]. Moreover, efficient human collaboration is often shown to be related to the degree that team members agree on, or are aware of task, role, and problem characteristics [13]. Thus, team mental model and SA are two precursors of team performance.

The efficiency of the team mental model includes assessing the similarity, overlap, and consistency of the individual mental models. For team SA, both environment and team dynamics need to be understood. However, each member does not have to be aware of every change; the common picture is shared by the team, not necessarily by all its members individually. As Gorman et al. discuss, better performance does not necessarily mean all team members sharing a common picture [14]. In addition, evaluating team cognitive precursors can also include evaluating workload distribution and social patterns and roles within the team.

3. GENERALIZABLE METRIC CLASSES

Based on the models presented in this paper, we can infer six generalizable metric classes relevant for human-robot team evaluation. Examples of sub-classes are included in brackets.

- Mission Effectiveness (e.g., key mission performance parameters)
- Human Behavior Efficiency (e.g., attention allocation efficiency, problem recognition efficiency, decision making efficiency, action implementation efficiency)
- Robot Behavior Efficiency (e.g., error-proneness, robustness, autonomy, learnability, memorability)

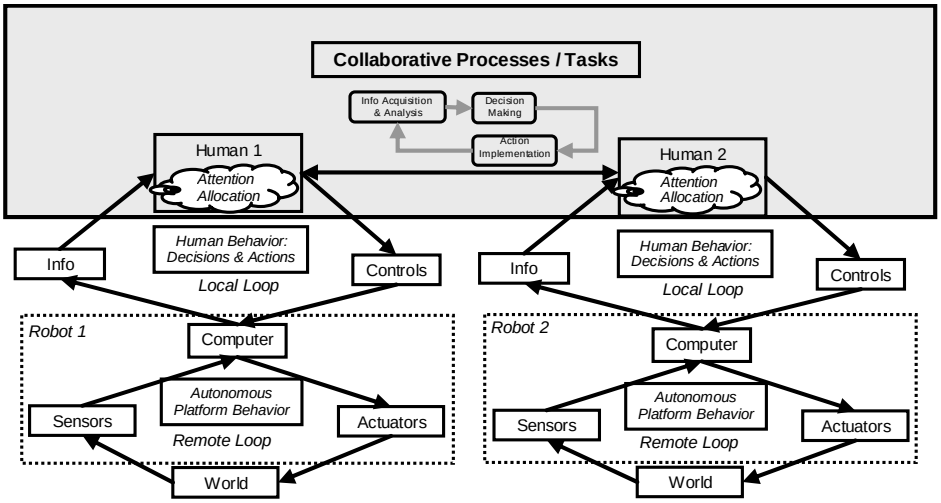


Figure 4. Human Collaboration in Supervisory Control of Robots.

- Human Behavior Cognitive Precursors (e.g., mental models, SA, mental workload, trust in automation, self-confidence, emotional state)
- Human Behavior Physiological Precursors (e.g., physical workload, physical comfort, physical fatigue)
- Collaborative Metrics

– Team Behavioral Action Efficiency (e.g., coordination efficiency, collaborative problem recognition efficiency, collaborative decision making efficiency, collaborative action implementation efficiency)

– Team Cognition Efficiency (e.g., team mental models, team SA, workload distribution, social patterns and roles)

– Robot Collaboration Efficiency

Evaluating the performance of the whole human-robot team requires applying metrics from each of these classes, but including metrics of every sub-class for every experiment can be inefficient and costly. As a rule of thumb, in addition to the more popular mission effectiveness and robot behavior efficiency metrics, incorporating at least one metric from the classes of human behavior efficiency, human behavior cognitive and physiological precursors, and collaborative metrics enables better team performance evaluation.

The next section discusses an experiment where a single human controlled multiple robots conducting a search and rescue mission. This study considered metrics for mission effectiveness, human behavior efficiency, and human behavior cognitive precursors. The value of incorporating metrics from each of these classes is discussed in the context of this experiment.

4. A CASE STUDY: SEARCH AND RESCUE MISSION

4.1 Experiment Description

In this experiment, a human participant teamed with multiple simulated robots to perform a search and rescue mission: removing objects from a maze[3] using different number of robots (2, 4, 6, or 8). The goal was a) to remove as many objects from the area as possible during an 8-minute session while b) ensuring that all robots were out of the maze when time expired. Collecting objects from the maze required the user to perform navigation and visual search tasks. First, the user assigned an object to the robot and the robot moved to that location. Second, the robot "picked up" the object, which in the experiment was simulated by the visual search of identifying a city on a map of the United States using Google Earth-style Software. Third, the user assigned one of the two maze exits to the robot and the robot carried the object out of the maze. The objects were randomly spread through the maze.

The maze was initially unknown, but the robots created and shared a map of the maze as they moved around it. Each robot could choose its path, choosing to explore an unknown path if it thought that path could possibly be shorter than the shortest known path to its user-specified destination. In addition, the robot would automatically choose an object or an exit after it had been waiting for a user-command for longer than 15 seconds. The user could at any moment redirect the robots to different locations by reassigning their destinations or rerouting them through a different path.

Sixteen people between the ages of 19 and 49 years old participated in the study. After completing a training and a comprehensive practice session, each subject participated in four 8-minute sessions, each with a different robot team size. The conditions of the study were randomized and counter-balanced. More details on the experimental setup can be found in [3].

4.2 Metrics Considered

This study measured metrics for mission effectiveness, human behavior, and human behavior cognitive precursors in an attempt to understand the final outcome of the mission, the decisions made and actions taken by the operator, and the causes driving those actions and decisions.

[3] In this experiment, the routes within the maze are unknown but the locations of objects to rescue are known.

We believe that at least one metric from each class is necessary to understand team performance. However, we recommend for the human behavior efficiency class, both attention allocation efficiency and human efficiency in conducting mission's tasks should be measured because they represent different aspects of the system. In addition, if the mission is composed of tasks of different cognitive nature, one human behavior efficiency metric for each task is also recommended. For the human behavior cognitive precursor class, the number of metrics selected depends on the actual research question and experimental setting. For this experiment, we measured trust and mental workload because both factors can influence human use of automation (i.e., robots' autonomy) [15]. Automation mistrust, which refers to over-reliance on automation, occurs in decision making because humans have a tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct [16]. This effect is known as automation bias.

We did not measure human behavioral actions separately for problem recognition, decision making, and action implementation because of the difficulty of distinguishing among these three categories in this particular testbed. No additional data that could support this analysis was recorded during the experiment.

This experiment did not measure metrics for human behavioral physiological precursors because with the 8-minute session time, these could not provide any meaningful insight. Collaborative metrics were also not considered since the focus was on single operator control, and robot efficiency was also not considered since they were simulated. Table 1 summarizes the metrics considered in this experiment.

Performance score, an indication of mission effectiveness, was defined as the total number of objects collected minus the number of robot lost (i.e., number of robots that did not get out of the maze when the 8-minute session expired).

HCIs were categorized in terms of robot navigation planning, robot navigation replanning, and visual search. The metrics selected were the time to complete a visual search, the time to assign a robot's destination, and the times to reroute a robot and reassign its destination.

The metric selected for attention allocation efficiency was the time required to decide which robot to service next, also known as the switching time. This metric included both the time it took for the user to decide which robot required his/her intervention, and the time required to select that robot on the display.

The frequency of overriding robot decisions was selected as an indication of operators' trust in robots. Finally, a five-point Likert scale was used to subjectively measure mental workload.

Table 1. Metrics Measured in the Case Study.

Metric Class	Selected Metric
Mission Effectiveness	• Performance score
Human Behavior Efficiency	• Average time to complete a visual search (indication of human efficiency in visual search) • Average time to complete a robot destination assignment (indication of human efficiency in planning robot navigation) • Average time to reroute a robot or reassign its destination (indication of human efficiency in replanning robot navigation) • Switching Time (indication of attention allocation efficiency)
Robot Behavior Efficiency	None
Human Behavior Cognitive Precursors	• Frequency of overriding robot decisions (indication of over-reliance on robots' autonomy) • Subjective rating of operator workload (indication of mental workload)
Human Behavior Physiological Precursors	None
Collaborative Metrics	None

4.3 Mission Effectiveness

Figure 5 shows the performance score as a function of the robot team size. A one-way ANOVA analysis showed that the robot team size significantly contributed to its variability (p-value = 0.018, R^2 = 15.31%). However, the R^2 of this model implies that it explained little of the performance variability. The Tukey test showed only difference in workload for 2 robots as compared to 8 robots.

Thus, evaluating performance in terms of robot team size does not provide much information, which confirms that additional metrics are required to really understand what happened in this experiment.

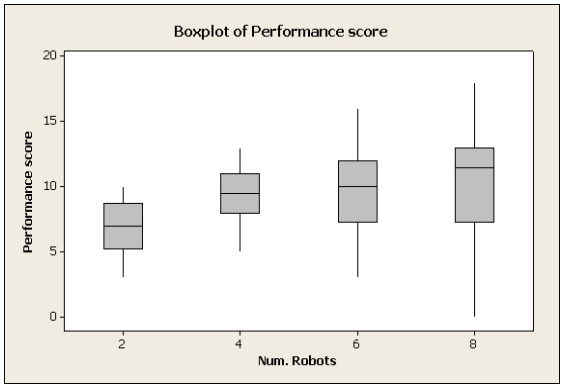


Figure 5. Performance Score vs. Size of the Robot Team.

4.4 Human Behavior Efficiency

Results suggest that the faster the subject completed a visual task, the higher the performance score (Pearson correlation = -0.594, p-value < 0.001). Results also suggest that subjects who were fast performing the visual search were also fast when selecting robot destinations (Pearson correlation = 0.479, p-value < 0.001).

Regarding navigation tasks, the average time to complete a destination assignment and that required to complete a reassignment are not correlated (Pearson correlation = 0.163,

p-value =0.214). This result confirms that the task of goal assignment for initial planning and for replanning were distinct.

Regarding replanning, robot destination reassignment ratio and rerouting ratio are strongly correlated (Pearson correlation = 0.526. p-value < 0.001), suggesting that subjects performed both reassignments and rerouting with a similar frequency. Results also suggest that people who were faster in the visual search, conducted more rerouting and reassignments (Pearson correlation of reassignment frequency & time for the visual search = -0.388, p-value = 0.002; Pearson correlation of rerouting frequency & time for the visual search = -0.345. p-value = 0.005).

Using an ANOVA model with the number of robots as the main factor and the average time to complete a visual search as a covariate, we obtained statistical significance for both variables (p-values < 0.001). The R^2 of this model was 59.38%, which means that 59.38% of the performance variability is explained with these two variables. The Tukey post hoc test showed only difference in performance for 2 robots as compared to the other robot levels. This result confirmed the trend seen in Fig.5 and additionally pointed that that there was also difference in performance for 2 robots as compared to 4 and 6 robots. Including in the ANOVA model other variables such as time to replan, or time to assign robot destinations did not improve the model. Thus, the average time to complete a visual search was the main factor driving the performance score. In this analysis, it was important to use these additional metrics to confirm our initial results and ensure consistency across metrics.

Regarding attention allocation efficiency, results show a strong correlation between performance score and switching time (Pearson correlation = -0.533, p-value < 0.001). Thus, performance scores tended to be higher with low switching times. Interestingly, the switching time and the time to complete a visual search are not correlated, which indicates that these are two independent sources of performance variability (Pearson correlation = -0.098, p-value = 0.441). This result demonstrates that the two human behavior metric classes (attention allocation efficiency and human efficiency in the visual search) are measuring different aspects of the system that should be considered separately to understand team performance.

4.5 Human Behavior Cognitive Precursors

Figure 6 shows that as the robot team size increased, subjects overrode fewer robot autonomous decisions. A one-way ANOVA analysis of the overriding frequency showed that the robot team size significantly contributed to its variability (p-value < 0.001, R^2 = 50.86%). The Tukey post hoc test showed only difference in overriding frequency for 2 robots as compared to the other robot levels. As task load, which refers to the task demands imposed on an operator, increased, users decreasingly overrode robot decisions. This result suggests that workload was affecting subjects' pattern for overriding automation.

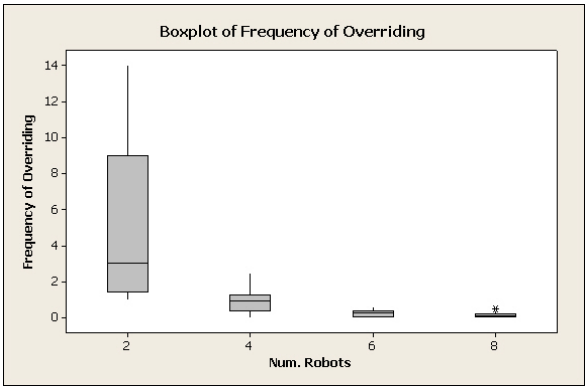


Figure 6. Overriding Robot Autonomy.

Additional investigation is needed to distinguish between subjects' cognitive saturation and subjects' over-reliance on robots. Subjective metrics for trust would allow further discussion. Since trust is a purely psychological state, subjective ratings are necessary to understand trust issues [17].

Figure 7 represents the perceived workload as reported by the subjects at the end of each scenario, 1 being nothing to do and 5 being completely overwhelmed. A one-way ANOVA analysis of workload showed that the robot team size significantly contributed to its variability (p-value = 0.005, R^2 = 18.86%). However, the R^2 of this model implies that it explained little of the workload variability. The Tukey test showed only difference in workload for 2 robots as compared to 6 and 8 robots. Subjective metrics are inexpensive and easy to administer, however they should be used to complement rather than to replace other forms of metrics.

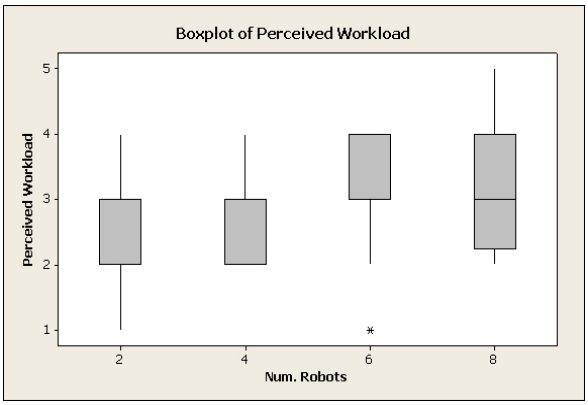


Figure 7. Perceived Workload.

4.6 Conclusions from the Case Study

This case study illustrates the need to measure multiple metrics across different metric classes to understand human-robot team performance, its underlying drivers, and effective design interventions.

In this experiment, analyzing human behavioral actions in the context of the tasks allowed us to identify that the visual search was the primary task driving the performance score. In addition, metrics of attention allocation efficiency pointed to an additional source of performance variability, switching time. Metrics of human behavior cognitive precursors allowed identifying that task load and over-reliance on robots' autonomy are interconnected. However, additional metrics for workload and trust that were not recorded during the experiment are necessary to distinguish between user cognitive overload and automation bias.

One potential drawback to the selection of metrics was that we did not explicitly measure behavioral actions in terms of problem recognition, decision making, and action implementation. Without this information, it is hard to say whether additional user support for problem recognition (e.g. which robot should I service next?) or decision making (i.e. which is the optimal route for this robot if I want to replan?) would be a better intervention to improve team performance. For interface design, measuring separately these three categories is essential because it allows exploring and understanding which parts of the mission require additional support and which design improvements can be more effective to maximize team performance. Measuring the complexity of the decisions that compose the mission and its workload as well as collecting more in-depth user feedback would also provide valuable information about future improvements.

However, problem recognition and decision making are highly interconnected and it can be difficult to measure them separately. As Klein and Klinger discuss, decision-making in complex environments under time pressure seems to be "induced by a starting point that involves recognitional matches that in turn evoke generation of the most likely action" [18]. Researchers should measure the observable outcomes of humans' decisions, and analyze and understand the decision process with other techniques such as verbal retrospective protocols.

5. CONCLUSIONS AND FUTURE WORK

This paper proposes a set of generalizable metric classes to consider for the evaluation of human-robot team performance. A case study of a single operator controlling multiple robots conducting a search and rescue mission illustrates the usefulness of measuring multiple metrics across these different classes.

Future work will populate these metric classes with the different types of metrics available and link them to actual research questions to help experimenters select the set of metrics that provide the most value for their experiments.

6. ACKNOWLEDGMENTS

This research was funded by MIT Lincoln Laboratory and U.S. Army Aberdeen Test Center.

7. REFERENCES

[1] Olsen, R., O. and Goodrich, M.A. 2003. Metrics for evaluating human-robot interations. In Proc. NIST Performance Metrics for Intelligent Systems Workshop.

[2] Steinfeld, A., et al. 2006. Common Metrics for Human-Robot Interaction. In Proceedings of the Conference on Human Robot Interaction (Salt Lake City, Utah, USA, March 2 - 3, 2006). HRI'06. ACM Press, New York, NY,

[3] Crandall, J.W. and Cummings, M.L. 2007. Identifying Predictive Metrics for Supervisory Control of Multiple Robots. IEEE Transactions on Robotics – Special Issue on Human-Robot Interaction, 23(5), 942-951.

[4] Sheridan T.B. 1992. Telerobotics, Automation, and Human Supervisory Control. The MIT Press. Cambridge, MA.

[5] Parasuraman, R., Sheridan, T.B., and Wickens, C.D. (2000). A model for types and levels of human interaction with automation. IEEE Transaction on Systems, Man, and Cybernetics--Part A: Systems and Humans, 30(3), 286-297.

[6] Wickens, C.D. and Hollands, J.G. (1992). Engineering psychology and human performance. Third Edition. New York: HarperCollins.

[7] Rouse, W.B. and Morris N.M. (1986). On looking into the black box: Prospects and limits in the search for mental models. Psychological Bulleting, 100, 349-363.

[8] Endsley, M.R. and Garland D.J. (Eds.) (2000) Situation Awareness Analysis and Measurement. Mahwah. NJ: Lawrence Erlbaum Associates.

[9] Drury, J.L., Scholtz, J., Yanco, H. (2003). Awareness in Human-Robot Interaction. In Proceedings of the IEEE Conference on Systems, Man, and Cybernetics, October 2003.

[10] Norman, D.A. (2002). The design of everyday things. New York: Basic Books.

[11] Durso, F.T., Rawson, K.A., Girotto, S.(2007). Comprehension and Situation Awareness. Handbook of Applied Cognition. Second Edition. Edited by Francis Durso.

[12] Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (in press). Advances in measuring team cognition. In E. Salas, S. M. Fiore, & J. A. Cannon-Bowers (Eds.), Team Cognition: Process and Performance at the Inter- and Intra-Individual Level. Washington, DC: American Psychological Association. (Orlando, Florida, USA, 26-30 September 2005).

[13] Fiore. S.M., Schooler J., W. (2004). Process Mapping and Shared Cognition: Teamwork and the Development of Shared Problem Models. In E. Salas, S. M. Fiore, & J. A. Cannon-Bowers (Eds.), Team Cognition: Understanding the Factors that Drive Process and Performance. Washington, DC: American Psychological Association.

[14] Gorman, J., Cooke, N., Pederson, H., Connor, O., DeJoode, J. (2005). Awareness of Situation by Teams (CAST): Measuring Team Situation Awareness of a Communication Glitch. In Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting.

[15] Parasuraman, R., Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. Human Factors 39, 230-253.

[16] Cummings, M. L. (2004). "Automation Bias in Intelligent Time Critical Decision Support Systems." Paper presented at the AIAA Intelligent Systems Conference.

[17] Wickens, C. D. and Xu, X. (2002). Automation Trust, Reliability and Attention HMI 02 03, AHFD-02-14/MAAD-02-2, AHDF Technical Report.

[18] Klein, G., & Klinger, D. (2000). Naturalistic Decision Making. Human Systems IAC GATEWAY, 11 (3), 16-19.

Toward Developing HRI Metrics For Teams: Pilot Testing In the Field

Jennifer Burke
jlburke4@cse.usf.edu

Kevin S. Pratt
kpratt@cse.usf.edu

Robin Murphy
Murphy@cse.usf.edu

Matt Lineberry
mlineber@mail.usf.edu

Meng Taing
mtaing@mail.usf.edu

Brian Day
bjday@cse.usf.edu

Center for Robot Assisted
Search and Rescue
University of South Florida
Tampa FL

ABSTRACT

This paper reports on the initial piloting of three instruments for studying human-robot teams: team member assessments of usability (effectiveness, ease of use, and satisfaction, and team compatibility), observer incident logs, and observer ratings of team processes. The pilot study was conducted during realtime human-robot operations at NIST's 4th series of rescue robot evaluation exercises, held in June 2007 at the TEEX "Disaster City" responder training facility in College Station, TX. In addition to the initial fielding of the General Robot Usability. Questionnaire and the Human-Robot Team Effectiveness Incident Log and Rating Scale with participating USAR Task Force members, we collected data consisting of video, field observations, and interviews. We tested techniques for gathering realtime data points on team processes and communication, and noted instances of emerging human-robot team work practices. Results for the initial pilot of the usability questionnaire (n=31 responses) yielded a reliability index of .94. The team process incident log was useful in capturing spontaneous multi-operator single-robot (MOSR) and multi-operator, multi-robot (MOMR) collaboration. Observers noted instances of all but one of the team process dimensions on the Rating Scale; however, contextual and experimental constraints prohibited a true pilot of the team process rating scale. The paper identifies needed revisions to these instruments. Once refined, these instruments will help in training human-robot teams, identifying best practices and techniques for human-robot team performance, and designing future robots that assist human teams in accomplishing increasingly difficult, dangerous tasks in critical, uncertain environments.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**] Robotics – *operator interfaces*.

J.4 [**Computer Applications**]: Social and Behavioral Sciences– *psychology*.

General Terms

Measurement, Experimentation, Human Factors

Keywords

Human-robot interaction, human-robot teams, team process, usability, field research methods, measures.

1. INTRODUCTION

Team operations in critical environments such as emergency response, military operations in urban terrain, and explosive ordnance disposal are beginning to incorporate robots and other remote presence technologies. As a result, new forms of distributed team processes are rapidly emerging and co-evolving with the accelerating improvements in technology. Understanding the relationships between robot activities and functions and human team member communications and coordination is key to successful human-robot teaming in critical environments. Standardized methods or techniques are needed to gather meaningful data, which is currently gathered ad hoc.

Developing human-robot team metrics, measures, and techniques is challenging for three reasons: no clear description of job tasks and activities exists; team roles for humans and robots are fluid, as are the protocols and procedures; and, in the case of rescuers, there is little variance in expertise across teams available for study as responders have little/no access to robot technology beyond the scope of these events. These issues are not endemic to robot technology, but commonly accompany the introduction of a radically new technology into an existing work practice. New work practices will emerge for existing task activities, and new task activities will be created as the new technology's potential is discovered.

This paper reports on the initial piloting of three instruments: team member assessments of usability (usefulness, ease of use, satisfaction, and team compatibility), and observer incident logs and ratings of team processes during realtime human-robot operations. The pilot study was conducted at the National Institute of Standards and Technology's (NIST) 4th series of rescue robot evaluation exercises, held in June 2007 at the Texas Engineering Extension Service (TEEX) "Disaster City" responder training facility in College Station, TX. Our purpose is to develop a toolset of standardized methods, metrics and techniques for measuring team processes and performance in human-robot teams. It is important to note that while we are employing the popular phraseology (referring to the human and robot as a team), we do not regard the robot as a team member, but as a resource used by the team. Human-robot systems currently used for real work are largely teleoperated, with low/no autonomy. Until the advances in autonomy/intelligence being made are manifested in fieldable robot systems, we shall

regard human-robot teams as teams of people using robots as a team resource.

Creation of standardized methods, metrics and techniques for investigating human-robot team processes will contribute to the development of training protocols for human-robot teams, and to the creation of performance evaluation standards for this new type of work. Measures created in this task can be applied to other efforts that are designed to improve team effectiveness through quantifying processes and are thus helpful in developing prescriptive teamwork models. The research is clearly relevant to coding schemes for activities in near real time. It is also relevant to the criterion problem, i.e. developing methods of identifying the best mixed team architectures.

2. RELATED WORK

Human-robot team metrics is an emerging topic, but current work does not yet adequately capture team processes of the human(s) and robot(s) working together or propose metrics which permit rapid feedback to evolving teams. Our work builds on our previous work in team processes by introducing a new coding instrument for recording team processes and investigating usability as a potential metric for team processes. Usability also been applied to human-robot interaction (HRI) but not specifically for team processes.

2.1 Human-robot Team Metrics

Metrics are a topic of great interest in the growing field of HRI, with many designing task or domain-specific measures [1]. Several attempts have been made to create a standardized set of metrics for HRI [2, 3]. None of these include team process metrics.

Pudenz et al. [1] offer an example of task/domain specific measures, as opposed to a standardized set. They considered techniques that quantified science team performance working with a remote rover robot, leading to an understanding of which features of the human-rover system were most effective and which features needed further development. Several of these variables were metrics and ratios related to the daily rover plan, the time spent programming the rover, the number of scientific statements made and the data returned. The most successful proxies for science effectiveness were the time to program each rover task and the number of scientific statements related to data delivered by the rover.

Yanco, Drury, and Scholtz [4] have used techniques similar to those presented in this paper to develop guidelines for designing interfaces for HRI. They used critical incident techniques, questionnaires, and interviews to analyze four different robot systems that competed in the 2002 American Association for Artificial Intelligence Robot Rescue Competition. They analyzed pre-evaluation questionnaires; videotapes of the robots, interfaces, and operators; maps of the robots' paths through the competition arena; post-evaluation debriefings; and critical incidents (e.g., when the robots damaged the test arena).

Our prior work focused on teams working with fieldable robots (air, ground, and water) in tasks such as urban search and rescue (USAR), explosive ordinance disposal, and mine rescue, but the metrics and methods developed to date are labor intensive. We have studied team processes using the Robot-Assisted Search and Rescue Coding Scheme (RASAR-CS) [5-7]. The RASAR-CS codes communications in terms of speaker/recipient, grammatical form, function, and content of the communication [8]. While the RASAR-CS is a valuable tool for studying HRI in team settings, it is time and labor-intensive, and can take many months to code a series of interactions. More timely metrics are needed. Burke [9] used onsite observers to rate team processes (communication, leadership, situation awareness, and backup) in high fidelity training exercises, and noted that other team process dimensions should be studied.

In the human teamwork literature, there is a vast body of research. In a review of the literature, Salas et. al [10] identified 138 different models of teamwork. In choosing an appropriate model for our purposes, three criteria were critical: 1st, that all components of teamwork be readily observable by assessors; 2nd, that the number of distinct components is manageable for assessors [11]; and 3rd, that assessments target behaviors that could be improved through training.

2.2 Usability

Usability testing, a traditional concept in human-computer interaction (HCI), can be used as a tool to determine whether a particular product or application is viable. If robots are to be used in team settings, we must ask whether they can effectively be used in a team context. Usability reflects the system's effectiveness, efficiency, and user satisfaction [12].

Several HRI studies have incorporated usability tests and measures, notably [4, 13, 14] but these have not addressed team processes. In a study comparing various levels of mixed initiative robot control, Bruemmer et al. [13] noted that the design of HRI and interfaces typically fails to follow basic usability principles or be informed by basic concepts of HCI. To address both these challenges, they used a development cycle of iterative usability testing and redesign to hone both the interface and the robot behaviors that supported it. Endo, MacKenzie and Arkin [14] used formal usability experiments to evaluate a mission-planning wizard into their MissionLab mission specification system, testing for usability improvements in terms of speed of the mission planning process, accuracy of the produced mission plans, and ease of use. Yanco and Drury [15] used usability testing, plus implicit and explicit situation awareness measurement techniques, to investigate USAR operators' levels of situation awareness and strategies for maintaining situation awareness.

2.3 NIST Rescue Robot Evaluations

As stated earlier, NIST is directly involved in creating appropriate metrics of usability for rescue robots. The robot assessments conducted by NIST include tests of visual acuity, mobility, directed perception, manipulator dexterity, and communications. Emergency responders work with robot developers in a set of deployment scenarios and NIST-generated tests to provide data for evaluation of the various robot platforms, and feedback on the methods used to generate the data. For humans-systems interaction they are developing standards related to criteria such as adequacy of initial training and proficiency education of operators provided by the robot developers. Especially relevant to this discussion is their treatment of "acceptable usability". They plan to use the percentage of tasks a user was able to perform without help as a measure of the effectiveness of the robot. Measuring efficiency would involve the time to completion for a task. As for user satisfaction, they suggest the use of a standardized satisfaction survey [16]. However, to the authors' knowledge, one has not yet been developed.

3. METHOD

This section presents both the General Robot Usability Questionnaire as well as the Human-Robot Team Incident Log and Effectiveness Rating Scale in addition to the field exercise and conditions under which they where they were evaluated.

3.1 Participants and Setting

The event was NIST's 4th Response Robot Evaluation Exercise, held at the TEEX "Disaster City" facility in College Station, TX from June 17-22, 2007. Disaster City is a first responder training facility designed to provide wide-area and specialty facilities which are difficult to arrange through other means. Among its various simulated scenarios, it has rubble piles, a multi-story office/parking structure collapse, a passenger train disaster, and a simulated hazardous materials train derailment. As for the Response Robot Evaluation Exercise itself, it was comprised of two distinct phases: first, a series of technical capability tests and second, a series of more realistic field scenarios. These initial tests were designed to test specific sub-components of the individual robot systems, and to evaluate NIST's response robot criteria they were developing. These tests included Aerial Station Keeping, Cache Packaging, Confined Space, Directed Perception, Mobility/Endurance, Grasping Dexterity, Inclined Plane, Random Maze, Radio Communications, Stairs, Steps, and Visual Acuity. The scenarios included a structural collapse and two train wreck sites. Several groups from the robotic and emergency response community attended the exercise. The most important group was the first responders who had made themselves available as sample operators to help pilot the various measures being tested. The measures being tested by NIST had originally been requested by DHS, thus the participants were from the FEMA USAR teams around the country; represented were Indiana-task force 1, California-task force 1, Colorado-task force 1, Maryland-task force 1, Nebraska-task force 1, New York-task force 1, Pennsylvania – task force 1, Texas – task force 1, Virginia – task force 1 , Virginia – task force 2, and Washington – task force 1. The robot platforms tested at the event were provided directly by the manufacturers, the majority of whom had onsite representatives to provide initial training and technical support. The final group was government and academic researchers comprised of researchers from NIST's Intelligent Systems Division of the Manufacturing Engineering Lab, and our team from the University of South Florida's Center for Robot-Assisted Search and Rescue. It should also be noted that TEEX is affiliated with Texas A&M Engineering. To the authors' knowledge, Texas A&M Engineering was not actively conducting research at the event.

3.2 Measures

3.2.1 General Robot Usability Questionnaire

One of the assumptions in designing the General Robot Usability Questionnaire is that the robots are best used in teams. Indeed, past research showed that when operators worked in teams during a training exercise, they were 9 times as likely to find simulated victims [6]. Therefore, the survey is divided into two sections. The first section asks about the user's opinion of the part of the robot operated by them. It consists of 4 items probing usefulness, 8 items about ease of use, 3 items asking about team compatibility, and 2 items concerning affective satisfaction. A list of the items, along with their respective dimensions for the first section is presented in Table 1.

Table 1. Questionnaire Items and Dimensions for Section 1

Item	Dimension
The device is effective for doing the task	Usefulness
The device's visual display tells me everything I need to know	Usefulness
The device allows me to do the task better than I could with other means	Usefulness
I am satisfied with what it can do	Satisfaction
I found using the device frustrating (reverse-scored)	Satisfaction
The device is easy to control	Ease of Use
The controls are designed logically and make sense	Ease of Use
The device's visual displays are easy to understand	Ease of Use
The device is prone to technical difficulties and malfunctions (reverse-scored)	Ease of Use
Learning how to use the device was easy.	Ease of Use
It is hard to make out what I'm seeing in the visual displays (reverse-scored)	Ease of Use
The design of the system makes it easy to understand what's going on with the device	Ease of Use
It is easy to get the device to do what I want it to	Ease of Use
During use, it is easy to know what my teammate(s) is (are) doing	Team Compatibility
During use, it's easy to communicate with my team member(s)	Team Compatibility
During use, it is easy to coordinate my actions with my team	Team Compatibility

Examples of these items include statements such as "The device is effective for doing the task", and "It is easy to get the device to do what I want it to do." The second part of the survey asks about the robot as a whole operated by the team. By nature, this section is more concerned with usability in a team context. It consists of 2 items probing usefulness, 4 items about ease of use, 4 items asking about team compatibility, and 2 items concerning affective satisfaction. Items included statements such as "The device is effective for accomplishing our mission", and "When using the device, it is hard for each team member to know what the others are doing." Items on both sections of the questionnaire are measured on a 1-5 Likert scale ranging from 1 (strongly disagree" to 5 "strongly agree." A "not applicable" category is included adjacent to the rating scale. A list of the items, along with their respective dimensions for the second section is presented in Table 2.

Table 2. Questionnaire Items and Dimensions for Section 2

Item	Dimension
The device is effective for doing the task	Usefulness
The device's visual display tells me everything I need to know	Usefulness
The device allows me to do the task better than I could with other means	Usefulness
I am satisfied with what it can do	Satisfaction
I found using the device frustrating (reverse-scored)	Satisfaction
The device is easy to control	Ease of Use
The controls are designed logically and make sense	Ease of Use
The device's visual displays are easy to understand	Ease of Use
The device is prone to technical difficulties and malfunctions (reverse-scored)	Ease of Use
Learning how to use the device was easy.	Ease of Use
It is hard to make out what I'm seeing in the visual displays (reverse-scored)	Ease of Use
The design of the system makes it easy to understand what's going on with the device	Ease of Use
It is easy to get the device to do what we want it to	Ease of Use
During use, it is easy to know what my teammate(s) is (are) doing	Team Compatibility
During use, it's easy to communicate with my team member(s)	Team Compatibility
During use, it is easy to coordinate my actions with my team	Team Compatibility

3.2.2 Human-Robot Team Incident Log and Effectiveness Rating Scale

In order to guide observations of teamwork in human-robot teams, we created a structured incident log and team effectiveness rating scale based on the team process taxonomy described by Dickinson and McIntyre [17]. In their taxonomy, teamwork is comprised of seven core components: backup behavior, communication, coordination, feedback, leadership, monitoring, and team orientation (see Figure 1).

In the model, *team orientation* refers to the interpersonal cohesiveness of the team, while *leadership* refers broadly to efforts to organize the team's actions, such as dividing roles and agreeing on a task strategy. *Monitoring* refers to teammates' awareness of one another's performance, which can lead to them offering *feedback* about each others' performance and/or offering to *backup* a teammate that is struggling. *Coordination* refers to the team's acting in concert with one another, which is dependent on their monitoring, feedback, and backup behavior. *Communication* among team members serves to link components, as when one team member monitors another's activities and uses communication to give feedback based on that monitoring. In our incident log, all of the above components are assessed except monitoring, which refers to teammate mental states that are very unlikely to be observable by an assessor.

For each team, an assessor kept a log of instances in which they observed the team demonstrating one of the teamwork components. Specifically, the assessor noted the time of occurrence, and then assigned a rating on a 1-5 scale of how well the behavior enhanced the team's effectiveness, with 1 representing behavior that significantly impaired team performance and 5 representing behavior that significantly enhanced team performance. Here, team performance refers to how well the team functioned together. It is considered separately from task performance, which is expected to be impacted by team performance but is also dependent on chance, individuals' skill, etc.

The assessor then specified who the agent of the behavior was, e.g. robot operator or mission specialist. Finally, the assessor noted which component of teamwork was demonstrated and what specific task the teamwork was directed towards, e.g. navigation, object identification, or search strategy. If two components were evident in a single incident, the assessor used their judgment to decide which component was most salient and assigned the incident to that component. Assessors are only asked to note incidents that they believe significantly impaired or enhanced the team's functioning. The intended outcome is a small set of significant incidents; as such, the assessor isn't overwhelmed with the task of logging every utterance made by a team member, and post-mission analysis and feedback can be focused and prompt.

Upon completion of each team's session, the assessor made an overall assessment of how effective the team was in each of the teamwork components, as well as their overall quality of teamwork and task work. These overall effectiveness ratings are not intended as a mathematical derivation from the number or quality of teamwork incidents observed during the session, but rather are based on the assessor's overall impressions of the team.

For both the incident log and rating scale, assessors were trained on general behavioral anchors for each process and each rating value, a copy of which is also available to them during

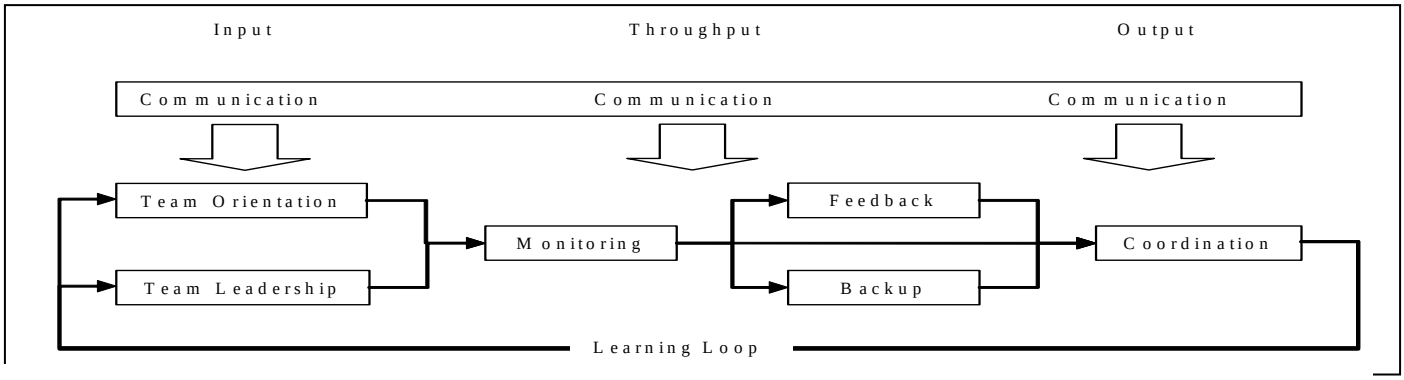


Figure 1. Dickinson and McIntyre's (1997) Team Process Model

the assessment. These descriptors were written to be easily understood and general enough to apply across different robot platforms and task domains.

3.3 Procedure

During the first two days of the event, the robots where cycled through the technical capability test (visual acuity, mobility, etc). The first day this was done by the vendors, and the second day, by the USAR responders. The first day allowed the vendors to become familiar with the tests and for the NIST teams to make sure the tests were set up appropriately. During the second day, robots were stationed at a given test and pairs of responders were cycled between the stations to give them training and exposure to all the various robots, and for the collection of a broader data set for a given robot/test pairing. During the second two days of the exercise the robots were stationed in small groups at one of the three scenarios, and the responder pairs rotated between the scenarios. The usability questionnaire was presented to the responders after each of the technical test sessions, as well as after each disaster scenario. As many of the technical tests were more to evaluate the robot hardware, rather than any team dynamics or coordination, portions of the usability questionnaire were not applicable in these case; the operators direct usability opinions of the system itself were still useful however. Occasionally we were unable to collect responses from individuals directly after a session; these missing responses were collected at the end of the event during the final hotwash (after-action meeting). As the incident log and rating scale was entirely focused on team interaction, the measure could only be employed during the larger scenarios. At each scenario one researcher took field notes and filled out the incident log, rating scale, and usability questionnaire at the end of each session. The second and third researcher on each team (we split into two teams of three) served as videographers for the session. One video taped the team itself and the second camera tracked the robots through the exercise (so that robot performance and operator mental models could be verified and correlated).

4. RESULTS

Results are reported from the piloting of the three instruments described in Section 3 during the NIST robot evaluation tests and scenarios. The initial pilot of the usability questionnaire (n=31 responses) yielded a reliability index of .94. The team process Incident log was useful in capturing spontaneous multi-operator single-robot (MOSR) and multi-operator, multi-robot (MOMR) collaboration. Observers noted instances of all but one of the team process dimensions on the Rating Scale; however, contextual and experimental constraints prohibited a true pilot of the team process rating scale.

4.1 General Robot Usability Questionnaire

The initial piloting of the usability questionnaire yielded mixed results, in that the usefulness, ease of use, and satisfaction items tested well, but the team compatibility items and second survey section were often skipped due to contextual and experimental constraints. The usability questionnaire (n=31 responses) yielded an acceptable index of reliability for the questionnaire overall (α = .94), usefulness (α = .85), satisfaction (α = .72), and ease of use (α = .87) dimensions. Note that these reliabilities do not include the second part of the survey or the team compatibility items, since contextual constraints prevented collection of an adequate amount of data for those items. The lower reliability index for satisfaction is likely due to having fewer items representative of that dimension.

We collected 31 questionnaires from 11 different responders using 10 different robot platforms in two of the NIST robot evaluation tests (manipulator dexterity and maze) and three responder scenarios (passenger train, HazMat train, and collapsed parking structure). Questionnaires were administered and collected immediately following the NIST robot evaluation tests, but this was not easily accomplished during the responder scenarios. The teams of responders typically worked at a scenario steadily until a hornblast sounded, signaling them to move to the next scenario and leaving them little time to complete the survey. (Not to mention the difficulty of wielding pencil and paper outside with few places to do so comfortably.) Instead, some responders completed the questionnaires during the afternoon hotwashes held late in the day, after the scenarios. This means there were time differences in when the questionnaires were filled out, and in the case of those completed after the scenarios, responders had in all likelihood used several different robots during the course of the day. They were asked to fill out separate questionnaires for each robot they used, but many felt they could only respond accurately about a few, saying that they did not have enough time with a particular robot to form an opinion. Furthermore, since the purpose of the exercises was primarily to give the responders a chance to practice using the robots, conditions were not controlled such as in an experiment. Thus, influence from developers varied greatly between trials. Therefore, the data from the scenarios is not presented here.

As stated earlier, one assumption underlying the survey design was that operators would be working in teams to operate one robot: this was not always the case. Data collection during manipulator dexterity training with the Telemax Unmanned Ground Vehicle (UGV) (n= 11) was typically from a single operator operating the robot. The Manipulator Dexterity Test was designed to measure the operator's ability to remotely grasp blocks and place them on different shelves of varying height. As with other methods, the main difficulty with using the test is that performance is both a function of the operator's ability and the capabilities of the robot. During the training, participants worked individually with the Telemax to practice grasping and moving objects. After having a chance to try out the Telemax, participants were asked to complete the usability questionnaire. However, since the robot was operated individually, participants did not fill out questions relating to using the device as a team. Therefore, no information was attained regarding the impact on team activities.

In general, the Telemax was rated as usable for the Manipulator Dexterity Test. The dimension scores of 4.1 for usefulness, 3.9 for ease of use, and 3.9 for satisfaction suggest that the participants agreed that the Telemax was usable. Note that the Telemax was chosen by NIST for training on the Manipulator Dexterity Test because it performed well when used by an expert operator. Thus, the general agreement between the usability ratings and objective performance on the task suggests that the items reflect usability. When examining individual items, it seems that responders thought it was effective and had useful, easy to understand visual displays. On the other hand, there was some disagreement about ease of operating it and ease of learning how to use it. However, it is important to note that the robot developer was present during the training (indeed, he conducted the training), and therefore responses may have been biased by this fact.

Our field observations of responders as they used the Telemax in this test suggest that operator experience, spatial ability, and technical knowledge may also influence the user's perceptions of usability. At a minimum, operator experience must be accounted for. Differences in the degree and type of interface feedback about the status and activity of the robot influenced responder perceptions of usability. For example, when operating the Matilda, the developer told the operator that a common mistake is that people forget that they are looking at the manipulator view. If the manipulator is not positioned straight and they think it's the main camera view, they will tend to drive crooked. Responders noted that a quad view or some other indicator of which camera view is being used would be helpful.

The nature of the task activity may also influence user perceptions of usability, in that differences may reflect whether the task is an existing one that could have been performed without a robot, or a new task that was previously not possible. In an instance that occurred with the Dragonrunner UGV at the Passenger Train scenario, the robot operator discovered that a wheel had fallen off only after another responder went into the train to locate a mannequin part. When searching for the part, he found the Dragonrunner's missing wheel. Team members treated this as they would a tool malfunction, i.e., it was not seen as a major flaw. A similar instance occurred with the Active Scope, a long, tubelike platform (shown in Fig. 2) that could go into very small voids previously inaccessible to responders. When using the robot, the operators thought they were making forward progress because they were able to insert more of the robot into the void. It turns out that the robot was simply coiling around in circles. Responders expressed misgivings about missing important information due to lack of feedback, but acknowledged the value of gaining a new method of access to confined spaces.

Figure 2. Operators use the Active Scope to explore a small void in the wall of a collapsed parking structure.

4.2 Human-Robot Team Effectiveness Incident Log and Rating Scale

As with the usability questionnaire, the results of the pilot tests for the incident log and rating scale yielded useful information in spite of the difficulties encountered in gathering data. The team process incident log was useful in capturing spontaneous multi-operator single-robot (MOSR) and multi-operator, multi-robot (MOMR) collaboration, but was physically unwieldy, involving multiple pages for a single scenario. Observers noted instances of all but one of the team process dimensions on the Rating Scale.

Conditions at this exercise were not ideal for observing teamwork between rescue personnel. Efforts to collect team process data during NIST's robot evaluation tests the first three days were unsuccessful. The robot developers were usually present, sitting with participants and guiding their actions or simply demonstrating the robot without allowing participants to operate them. The work domain of search and rescue is very difficult, dangerous, and stressful, as rescue personnel race against the clock in unstable and low-visibility environments to find trapped victims. Such work demands teamwork for safety, emotional support, and maximum performance. However, the scenarios in this exercise were relatively simple and involved none of the dangers or time pressures inherent to search and rescue. Further, many participants had not been trained to proficiency on the robots, so much of the time in scenarios was dedicated to basic training on robot operation. This is not an exception to the rule; despite the fact that the responder community, government, and public all see robot technology's potential value in future response operations, this community (emergency response) does not have the sustained access to these technologies necessary to train and practice to an acceptable level of proficiency.

While teamwork was not encouraged or facilitated at this exercise, we nonetheless observed incidents of teamwork with both multiple operator-single robot (MOSR) teams and multiple operator-multiple robot (MOMR) teams. One factor that may have influenced the emergence of spontaneous teamwork is the amount of physical coordination needed to operate a given robot. For example, the Active Scope robot resembles a rope and has a camera at the front end. The front end of the robot's movements can be controlled using the operator control unit (OCU). However, the back end of the robot needs to be manipulated manually. As such, teams tended to have one person observing and operating the OCU while another person would manually manipulate the back end of the robot. The physical coordination and cooperation needed to operate the robot stimulated the frequency of communication between the team members. When using the Active Scope, there was an instance when it was getting hung up. The person observing the OCU was trying to explain to the team member manually operating the back end what he was seeing. The team member manually operating the back end wanted to be sure he understood correctly, so he found a rock on the floor and etched a picture on the concrete to confirm his understanding of the OCU operator's description. This instance provided clear examples of backup, communication, and coordination in the incident log for that session. It also demonstrates how a second view for the manual operator would have been effective in helping the team establish common ground. An example of the multiple operator-multiple robot paradigm was observed when teams decided to deploy a Dragonrunner along with a Packbot UGV to search the inside of a passenger train for victims. The Dragonrunner robot operator offered to act as a spotter for the Packbot operator and his teammate, providing examples of leadership and backup behavior. Additionally, the Dragonrunner operator gave the Packbot operator feedback about the thoroughness of his search, noting that the Packbot had missed a searchable void. The Packbot operator acknowledged his teammate and adjusted his search accordingly (coordination). Without the teammate's help, it is doubtful that the Packbot operator would have conducted a thorough search of the mission space. Regarding the incident, the Dragonrunner operator remarked, "Redundancy: it's a beautiful thing."

In another multiple operator-multiple robot scenario, a participant controlled a Matilda UGV to investigate a HazMat train wreck while another participant (AirRobot operator) used

a Heads-Up Display (HUD) to direct and receive feedback from an AirRobot unmanned air vehicle (UAV), which was teleoperated by an AirRobot vendor acting as UAV pilot. The AirRobot operator offered to share his view with the Matilda operator, who had become lost; in the team process taxonomy, this would be classified as coordination and backup behavior. The following transcription captures a moment of emergent teaming:

Airbot Operator: (offering HUD to Matilda Op.) Here, do you want to see this video at the same time?

Matilda Operator: (laughing) Don't screw me up.

(general laughter)

Airbot Pilot: Ok, but I can see you down there. I can hear you crawling.

Airbot Operator: There, you're going. You're following the train car.

It would appear that the Matilda operator rejects the offer because he is already in visual information overload (he was lost), and can't watch any more video. Realizing this, the AirRobot Operator puts the HUD back on and instead begins to talk his teammate through localizing himself (and they keep up the auditory collaboration through much of the session).

5. CONCLUSIONS/ RECOMMENDATIONS

A set of three instruments designed to capture human-robot team processes and team member assessments of usability (effectiveness, ease of use, satisfaction, and team compatibility) was piloted at the NIST Robot Evaluation Exercise held in June, 2007 in “Disaster City”, a high fidelity emergency response training facility in College Station, TX. The primary finding was that operators worked in teams of 2-3 to perform robot-related task activities, even when (or in spite of) circumstances and contextual conditions were oriented toward single operator-single robot configurations, as was the case for many of NIST’s robot evaluation tests. We were not able to document enough instances of team processes to report meaningful findings on the team process instrument, but the observations/experiences collecting the data provided important information for improving techniques to study human-robot teams.

Initial administration of the instruments revealed several factors to be considered during revision:

• Paper instruments are unwieldy and difficult to complete immediately after a robot operation due to the nature of the context.

• The questionnaire does not account for operator experience, different team configurations (SOSR, MOMR) and level of coordination required.

• Responses may differ depending on whether the team is performing a task that has previously be done without robot technology, or a new task/activity that using the robot has afforded.

• Particularly for brief scenarios, not all teamwork categories will be apparent. Rather than asking the rater to give overall ratings on such categories, a "not observed" option needs to be available.

Figure 3. Responder (AirRobot Operator) wearing the AirRobot’s Heads-Up display offers to spot for the Matilda robot operator (seated to the left). The AirRobot pilot (far right) watches the OCU display as others observe.

In addition to the revisions noted above, we plan to transition the instruments piloted in this report to a tablet PC format. One benefit of a tablet PC format is that it could allow the rater to make notes "on the fly" about infrequent but relevant occurrences that do not fit in the team process scale, such as the addition of a new team member. Raters could fluidly select from different rating modes to create a very descriptive narrative.

Based on the lessons learned from these proto-studies, we conclude that the toolkit of methods and techniques currently under development requires a dedicated physical testbed where experimental and contextual factors can be controlled and standardized. Moreover, access to teams that use robots as part of their normal work practices is needed to assure variance in experience levels and performance. This suggests studying military and civil explosive ordinance disposal teams, currently the only such user group, to inform the development of effective USAR human-robot teams. To understand and promote the most effective human robot team work practices, methods for both MOSR and MOMR must be explicitly studied and developed. These methods and techniques will 1) provide quick feedback to team members; 2) give leaders objective information on how to best leverage robot technology in the workplace (e.g., effects upon existing team work practices, what new practices emerge) based upon valid, reliable measures; and 3) serve as a source of relevant user data to researchers, robot designers, and the HRI community.

These outcomes will help human-robot teams in critical environments such as natural or manmade disasters function more effectively, thus enabling them to reduce loss of life in these events. It will also help human-robot teams in less critical environments by shedding light on the challenges posed by new technology insertion, and the models/procedures that emerge to effectively incorporate new technologies into distributed team systems.

6. ACKNOWLEDGMENTS

This work was sponsored in part by the US Army Research Laboratory under Cooperative Agreement W911NF-06-2-0041. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARL or the

US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. Our thanks to NIST, especially Elena Messina and Adam Jacoff; Billy Parker and the staff at TEEX, and the responders and robot developers who graciously allowed us to get in their way.

7. REFERENCES

[1] E. Pudenz, G. Thomas, J. Glasgow, P. Coppin, D.Wettergreen, and N. Cabrol, "Searching for a quantitative proxy for rover science effectiveness," in 1st ACM SIGCHI/SIGART conference on Human-robot interaction, Salt Lake City, Utah, USA, 2006, pp. 18-25.

[2] J. W. Crandall and M. L. Cummings, "Developing performance metrics for the supervisory control of multiple robots," in ACM/IEEE international conference on Humanrobot interaction, Arlington, Virginia, USA, 2007, pp. 33-40.

[3] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common metrics for humanrobot interaction," in 1st ACM SIGCHI/SIGART conference on Human-robot interaction, Salt Lake City, Utah, USA, 2006, pp. 33-40.

[4] H. A. Yanco, J. L. Drury, and J. Scholtz, "Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition," Human-Computer Interaction, vol. 19, pp. 117-149, 2004.

[5] J. Burke and R. Murphy, "RSVP: An investigation of remote shared visual presence as common ground for human-robot teams," in ACM/IEEE 2nd International Conference on Human-Robot Interaction Washington, DC, 2007.

[6] J. L. Burke and R. R. Murphy, "Human-Robot Interaction in USAR Technical Search: Two Heads are Better Than One," in IEEE RO-MAN 13th International Workshop on Robot and Human Interactive Communication, Kurashiki, Okayama, JAPAN, 2004.

[7] J. L. Burke, R. R. Murphy, M. D. Coovert, and D. L. Riddle, "Moonlight in Miami: A field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise," Human-Computer Interaction, vol. 19, pp. 85-116, 2004.

[8] J. L. Burke, R. R. Murphy, D. R. Riddle, and T. Fincannon, "Task Performance Metrics in Human-Robot Interaction: Taking a Systems Approach," in Performance Metrics for Intelligent Systems, Gaithersburg, MD, 2004.

[9] J. Burke, "RSVP: An Investigation of the Effects of Remote Shared Visual Presence on Team Process and Performance in Urban Search & Rescue Teams," in Department of Psychology. vol. Ph.D. Tampa, FL: University of South Florida, 2006.

[10] E. Salas, K. C. Stagl, C. S. Burke, and G. F. Goodwin, “Fostering team effectiveness in organizations: Toward an integrative theoretical framework of team performance,” in Modeling complex systems: Motivation, cognition, and social processes. W. Spaulding and J. Flowers, Eds. Lincoln, University of Nebraska Press, 2004.

[11] B. B. Gaugler and G. C. Thornton, “Number of assessment center dimensions as a determinant of assessor accuracy. Journal of Applied Psychology, vol 74, pp. 611-618, 1989.

[12] J. Rubin, Handbook of usability testing: How to plan, design, and conduct effective tests. New York: Wiley, 1994.

[13] D. J. Bruemmer, R. L. Boring, D. A. Few, J. L. Marble, and M. C. Walton, ""I call shotgun!": an evaluation of mixed initiative control for novice users of a search and rescue robot," in Systems, Man and Cybernetics, 2004 IEEE International Conference on, 2004, pp. 2847-2852.

[14] Y. Endo, D. C. MacKenzie, and R. C. Arkin, "Usability evaluation of high-level user assistance for robot mission specification," Systems, Man and Cybernetics, Part C, IEEE Transactions on, vol. 34, pp. 168-180, 2004.

[15] H. A. Yanco and J. Drury, ""Where am I?" Acquiring situation awareness using a remote robot platform," in Systems, Man and Cybernetics, 2004 IEEE International Conference on, 2004, pp. 2835-2840 vol.3.

[16] E. Messina, A. Jacoff, J. Scholtz, C. Schlenoff, H. Huang, A. Lytle, and J. Blitch Statement of requirements for urban search and rescue performance standards, 2005. Preliminary version. (http://www.isd.mel.nist.gov)

[17] T. L. Dickinson and R. M. McIntyre, " A conceptual framework for teamwork measurement," in Team Performance Assessment and Measurement: Theory, methods and applications, M. T. Brannick, C. W. Prince, and E. Salas, Eds. Mahwah, NJ: Erlbaum, 1997

Framing and Evaluating Human-Robot Interactions

Curtis W. Nielsen, David J. Bruemmer, Douglas A. Few, David I. Gertman
Idaho National Laboratory
P.O. Box 1625
Idaho Falls, ID, USA
{Curtis.Nielsen, David.Bruemmer, Douglas.Few, David.Gertman}@inl.gov

ABSTRACT
The field of human-robot interaction (HRI) has been around in some for or another for some time. Recently there has been a push for systematic analysis methods and metrics that can be used to evaluate research in HRI. The challenge with developing metrics for the field of HRI is that the field is so broad---ranging from social robots to toy robots, from military robots to therapeutic robots with different areas of importance for each type of robot. Prior proposals of taxonomies and metrics have resulted in well defined but narrowly scoped solutions. The purpose of this paper is to provide a simplified framework that allows the general classification of different human-robot interaction experiments such that appropriate comparisons can be made between experiments from different researchers. Furthermore, we discuss two common metric classes: evaluations of the task objective and human involvement, and the often overlooked metric class of reliability measures which can all be used to categorize the results from various HRI experiments and provide insight into what needs to be done with the field of HRI as a whole.

Categories and Subject Descriptors
I.2.9 [**Artificial Intelligence**]: Robotics – *Autonomous vehicles, operator interfaces, sensors.*

General Terms
Algorithms, Measurement, Performance, Design, Reliability, Experimentation, Human Factors, Standardization, Verification.

Keywords
Framework, evaluation, human-robot interactions, performance, metrics, reliability.

1. INTRODUCTION
The field of human-robot interaction has been around in some form or another for some time [7]. Most of the research to date has been performed by independent research groups who develop their own set of metrics and measures of performance for evaluating the success of the robot system. Only recently has there been a real impetus for systematic analysis methods and tools that can be applied to robotic systems and used across domains.

The challenge with attempting to design metrics for human-robot interaction, is that the field is so broad, making it difficult for any one set of metrics to support generalized comparisons across domains. For instance, the field of HRI involves interactions with technology ranging from "toy" robots to military robots, and from humanoid personal assistant robots to vacuum cleaning robots, or from robots that comfort people to those that are designed to keep them out of harms' way. Each of these robot systems will have different performance metrics for evaluating their usefulness, helpfulness, playfulness, inspiration, entertainment value, etc. Moreover, while some robot systems may rate well with one set of evaluations, they may not rate well with another yet still be valuable to the purpose for which it was designed.

The challenge is to develop a methodology for considering human-robot systems that supports, rather than restricts the broad development of human-robot interactions and robot technologies, but also allows a means for evaluating research and making meaningful comparisons between the different research approaches and solutions over such a broad field. While there have been proposals for taxonomies [32] and metrics [28] for HRI, these are often cumbersome and difficult to implement because they address the details of a particular area of research and are generally not flexible to encompass a larger variety of research and conclusions. In contrast, we are looking for a framework that is simplified and supports a general understanding of the differences between experiments without defining the nitty-gritty that is so often institution and experiment-specific.

At the Idaho National Laboratory (INL) our human-robot interaction research has focused primarily on the use of robots as tools that can effectively help first responders, soldiers, and other individuals involved in emergency response and other critical, hazardous endeavors. To develop the technologies correctly required numerous user studies of differing types and situations including simulation, the physical world, different environments, different tasks, and participants with varying levels of prior knowledge regarding the task and use of robots. All the different experiments were required because each led to new insights about how people view and use robots.

For instance, some experiments involved participants who were complete novices at the domain of interest and use of robots, while others knew the domain but had never used robots and still others had used robots to accomplish tasks within their domain of expertise. Without such a wide variety of experiments, it would be difficult to develop a robust and reliable system that had sufficient capabilities for an end-user but also necessary simplicity such that domain-users could apply the robot to their domain without the requirement of understanding all the nuances of the robot system and, in effect, being a robot expert.

While it is understood that not all research institutions are going to be able to systematically answer every question related to the development of human-robot interactions, it is beneficial to provide a framework where the contributions of one group can benefit other groups and as a community the complete picture can be established.

Therefore, the purpose of this work is to identify relevant issues and associated questions that can be used to address the categorization of human-robot interaction evaluations which can then be used for framing correct comparisons. To that end, this paper will present a loose framework that can be used to frame human-robot interaction evaluations and measurements into simplified categories based on the purpose of the robot, what is being evaluated, and the prior skill of the participants.

We will then illustrate how previous human-robot interaction experiments fit into the framework of experiment categorization. Following this discussion we will review the measures of performance from the experiments and categorize them into two common metric classes that measure human involvement and task performance and we will discuss the often overlooked metric class of reliability.

2. FRAMING THE EVALUATION OF HUMAN-ROBOT INTERACTIONS

One of the challenges with the field of HRI is that the research area is so broad that it is difficult to define metrics or taxonomies that are relevant for a broad area, yet specific enough to make meaningful comparisons between different research solutions from different institutions.

In order to properly frame the evaluation of human-robot interactions, it is important that we ask the right questions regarding the purpose for which the robot was designed, built, and tested. To accomplish this, we have identified three general areas that should be addressed when discussing human-robot interaction evaluations: What is the overall objective of the robot in the experiment? What are the types of participants that can be used to evaluate the system? And, what, about the system, is being evaluated?

2.1 What is the overall objective of the robot?

The purpose of this question is to separate experiments that are focused on the social aspects of human-robot interaction (e.g. museum tour guides, playmates for children, therapeutic assistants) with those that are used more as tools to accomplish tasks that could be dull, dirty, or dangerous to humans so that adequate comparisons within the categories of social robots and "tool-based" robots can be made.

2.1.1 Social interaction

Cynthia Breazeal provides an excellent discussion on what it means to be a "social robot" [3]. She defines a social robot as "[one] that people apply a social model to in order to interact with and to understand." She also notes that social models are generally applied by humans to explain, understand, and predict the behavior of complex non-living things that act autonomously (also see [25]). Indeed the term "social" has changed over time to become more associated with anthropomorphic social behavior [3], [12]). Breazeal goes on to point out a few characteristics of social robots including:

- *Socially Evocative:* Encourages people to anthropomorphize the technology in order to interact with it.
- *Socially Communicative*: Uses human-like social cues and communications to facilitate interactions with people.
- *Socially Responsive:* Perceives human social cues and can benefit and learn from people.
- *Sociable:* Socially participative robots with their own internal goals and motivations.

A very brief set of examples of work with social robots include Kismet [2], a wedding photographer [8] , robots in classrooms [30] museum tour guides [31] and therapeutic assistants [17].

2.1.2 Tool

One definition of a robot that is used as a tool is one that is used to perform tasks that are dull, dirty, or dangerous to humans. In general, the operator does not really care about the robot as an individual rather he or she just wants to get the job done. While humans may attribute anthropomorphic qualities to robots that are used as tools, the robot is not necessarily designed to support the qualities of social interaction. Some examples of work with robots as tools includes search and rescue [20], SWAT teams [16], military [19], and space exploration [1].

It should also be noted that sometimes a robot solution will cross the boundaries between the social and tool classification, especially in domains where the robot is used in a challenging environment, but as part of a team requiring adequate social models to understand what each of the team members are doing and where the information relies. Some examples of research where the experiments could have both tool and social elements include the integration of robots into a search and rescue task force [6].

2.2 What is being evaluated?

Once we know the overall objective of the robot, we can then focus on what the experiment is evaluating. Experiments are generally designed to answer some hypothesis regarding the usefulness of different conditions of the interaction, whether it is the algorithms, the interface, control mechanisms, environment, robot, or communications.

One of the biggest challenges with human-robot interactions is that there are so many variables that affect the human-robot system that it quickly becomes intractable to empirically test all combinations and tease out true differences between solutions. When we review the literature, we also see that many experiments are evaluating different aspects of the system, so it is difficult to determine if one experiment can or should be compared to another experiment. To classify experiments with respect to what is being evaluated, we present five categories of evaluations: algorithm; component; sets of components; system; and sets of systems.

2.2.1 Algorithm

Algorithm evaluation is used to prove that a particular and fundamental algorithm is working correctly and to gain a better appreciation of the limitations and capabilities of the algorithm. In order to make effective human-robot interactions, there are countless possibilities of algorithms that could be evaluated, tested, and proven. Moreover, an appropriate understanding of the limitations of the algorithm will enable the operator to know where they should focus their efforts to improve the system. Some examples of research that might fit in this category include algorithms for feature extraction, path-planning, information visualization, or obstacle avoidance.

2.2.2 Component

The component level of evaluation includes tests that compare the value of multiple algorithms combined into a solution for a larger problem such as how information should be displayed on the interface, which robot behavior works best, or which interaction mode is most efficient. To correctly design an experiment evaluating components of the human-robot system. All aspects of the system except the part in question should remain constant. For example, if we were considering different interface designs (positioning of map and video or 2D interface vs. 3D interface), we should keep the robot behaviors (e.g. guarded motion), robot, environment, and interaction mode (e.g. joystick) the same. If we were comparing different robot behaviors, we should keep the robot, the interface, the environment and the command interface the same while only modifying the robot behaviors.

2.2.3 Set of Components

In human-robot interactions, it is simply unrealistic to evaluate all the different component comparisons because there are so many different components related to the full system. What adds more to this challenge is that sometimes individual components are related to each other in terms of performance and usability. For example, one particular interface design might be better than others for one set of robot behaviors, but for another set of behaviors, a different interface design might be better.

While some may criticize experiments that modify multiple aspects of the human-robot interaction, if the changes are done systematically and the experiment is designed appropriately, we can still gleam valuable results. For these types of experiments, the things that remain constant include the robot itself, the communications from the robot, and the environment. One of the limitations of these experiments is that it would be unwise for the experimenters to make claims regarding the efficacy of an individual component because the effects of individual components are affected by the full set of chosen components.

2.2.4 System

While evaluating sets of components provides insights that may lead to a best case solution for a single robot system it does not provide the framework for comparing one complete robot system against an entirely different robot system. For that reason, the next set of experiments evaluates complete HRI systems which include the robot, behaviors, communications, interface, and control capabilities. An example of this approach would be to compare the iRobot PackBot with the Foster-Miller Talon where each robot is operated with its own operator control unit and its own communications. Proper experiment design would entail that while nearly everything about the robots may be different, the task and the environment should remain the same. This type of experiment is particularly valuable at determining which robot system is best for certain tasks and environments as well as determining the limitations of systems. It is also useful to determine if "innovations" have actually improved the state-of-the-art.

2.2.5 Set of Robot Systems

Once a full system has been evaluated, the next set of experiments involves the use of multiple systems. While that may (or may not) seem like a simple step forward, this set of experiments includes a large variety of experiments that can include one or more operators and one or more robots of potentially differing capabilities. Yanco and Drury discussed a taxonomy related to the various possibilities of multiple-human multiple-robot interactions [32]. For the purpose of framing experiments it suffices that the experiment could be classified as evaluating a set of systems if there is more than one robotic system in place. Others have worked on how to address the evaluations of multiple agent systems from an HRI context [10, 23].

2.3 Who are the participants?

With an understanding of the purpose of the robot and what exactly is being evaluated, the next set of classifications is based on the types of participants used for an experiment.

Pedahazur and Schmelkin describe many of the pitfalls in research related to selection and handling of study participants [24]. Their admonishment to select the right subjects is essential because novices and experts do not perform the same [15]. When evaluating reports on human-robot system, participants come from a varied set of backgrounds and experience with robots or other related technology such as remote control cars or airplanes. The purpose of this section is to help differentiate between the types of participants that may be involved in experiments. The very general classification of experiments can be broken down into two primary groups: those with and without participants. For the category of those with participants, the characterization of participants is often separated broadly into experiments with novices (e.g. students, the public) and experts (e.g USAR, Military). However, as robots are introduced into fields where personnel have different levels of training with robots as applied to their domain, we again come to the question of novice versus expert, but this time with respect to the end-user's experience with robotics.

TABLE 1. User Group Experience

	No Robot Experience	**Robot Experience**
No Domain Experience	Students The public	Engineers Developers
Domain Experience	SME-General SME-Specific	SME- Specific.

Experimental design and debriefing must carefully distinguish between the various kinds of end users and the different levels of user experience. All so called "subject area experts" are not created equally. In fact, in a recent study with radiological hazard detection, three different groups of subjects were involved including personnel with robot operation and dirty bomb response training; those with only dirty bomb response training and those subjects with general training with radiation detection. In the experiment, treating these users as if they were all the same would have been an unfortunate mistake since certain features of the robot behaviors and interface were used very differently and, in fact, during analysis user experience turned out to be a significant factor. When evaluating robotic systems there are two types of experts that should be addressed: domain experts and robot experts. The types of participants with the varying levels of expertise are shown in Table 1 and discussed below. The table and some of the following discussion were presented previously [21].

It is understood that for some domains (e.g. social robotics) it may be difficult to distinguish between participants that do and do not have domain expertise. In these cases, there may be differences between people who have used the robot previously and have some sense of the robot and how to interact with it, and those who have not interacted with a robot previously. These differences in participants should be noted in experiment reports.

In situations that do require domain knowledge such as search and rescue or explosive ordinance disposal, it is important to understand the domain expertise of the participants and how their robot expertise helps clarify how their data should be used and interpreted. The following discussion relates to the four combinations of robot and domain expertise. We present the discussion in progressive order from the "Robot-No Domain" category counter-clockwise to the "Robot-Domain" category. This also represents the order of experiments that we believe leads to technologies and experiments that are best suited to apply to domain experts.

2.3.1 Robot – No Domain

Members in this group include the robot developers and engineers who are intimate with the workings of the robot and capabilities, both realized and potential, but not necessarily knowledgeable with the domain or how robots should be applied to the domain. This group of users could be considered an early pilot study group because their efforts are to make the robot work how they think it should and to respond to requests and suggestions from participants in other experiments.

2.3.2 No Robot – No Domain

Members in this group are those unfamiliar with the domain and who have not really used robots before. This group could include students and the general public and is particularly valuable for evaluating the core functionality of the robot that is not dependent on knowledge of the domain. For example, users in this group are beneficial to test levels of autonomy, interface designs, different control schemes, as well as different robot systems, all with the goal of understanding how the robot can and should be used in general, but not necessarily specific to any domain. Experiments with participants in this group should include tasks that are similar to those that might be performed with SMEs, for the purpose of discovering the general principles of human-robot interaction that may apply to particular domains.

2.3.3 No Robot – Domain

Members in this group are those who have been trained in a specific domain and could be considered subject matter experts (SMEs), but who have not used robots as applied to their domain. This group could further be divided into a variety of expertise levels depending on experience and training particular to the domain however, in general, it suffices to say that SME participants either have general training regarding issues relevant to the domain, or specific training regarding tasks within the domain. As an example, a nuclear engineer might have generic radiation training and a Civil Support Team member might have general radiation training as well as emergency response training specific to radiation hazards.

This group is valuable for beginning the discussions about how domain users will approach the use of the robot for a task that they are familiar with. Furthermore, as robots are introduced into fields where they are not prominently used, these experiments support a collaboration of ideas between researcher and SME in that the SME can provide insights to the robotics researcher about how they approach their task and what their concerns are while the robotics research can help the SME understand the limitations and potential of the technology.

2.3.4 Robot – Domain

Members in this group are similar to the SMEs in the previous group with the additional requirement that they have had experience or at least training using robots within their domain of expertise. As an example, a participant in this category might be an explosive ordinance disposal (EOD) trained individual with experience using a robot to accomplish the EOD mission. Members in this category are helpful at evaluating new innovations against the current state of practice.

3. CLASSIFICATION OF PREVIOUS HRI EXPERIMENTS

The purpose of this section is to use the aforementioned framework and characterize previous experiments with robots according to the proposed framework. For each experiment we briefly state the goal, what the robot is used for, what is being evaluated, the participants, dependent measures, and lessons learned. Most of the experiments are focused on robots that are used as tools. Future work would identify the classifications of experiments involving social robots.

3.1 Interface design [22]

Purpose: Compare 2D and 3D interface solutions for mobile robot teleoperation task using a joystick and a pan-tilt-zoom camera.

Social robot or tool? Tool.

Test: Component (interface design).

Participants: (18-176) No Robot – No Domain (Students).

Dependent Measures: Task completion time, collisions, NASA-TLX workload, Behavioral entropy workload, proximity to obstacles, average velocity, distance traveled, objects found, time to identify objects.

Lesson Learned: Operators performed better with the 3D-egocentric interface with respect to completion time, collisions, workload, and proximity to obstacles. Operators were able to find objects faster when using the 3D interface, but most likely because it was easier to navigate with the 3D-egocentric interface, not necessarily find things.

3.2 US Army XUV [14]

Purpose: Examine use of scalable interfaces (tablet vs. screen) and operator span of control with UGVs performing autonomous mobility and Reconnaissance, surveillance, and target acquisition.

Social robot or tool? Tool.

Test: Component.

Participants: (2) No Robot – Domain (Soldiers).

Dependent Measures: Span of Control. Distance Traveled, number of teleoperations, number of back-ups, number or RSTA images sent, mission time, verbal spot reports, workload (NASA-TLX).

Lessons Learned: Interface size did not have major effect, although participants preferred smaller tablet size. Operators appeared to be able to manage two robots (including monitoring, RSTA reports, and operator interventions).

3.3 Robot behavior development [4, 5]

Purpose: Compare different modes of autonomy and interface design for indoor search and exploration tasks.

Social robot or tool? Tool.

Test: Component, Set of components.

Participants: 19--120 No Robot – No Domain (General public).

Dependent Measures: number of objects found, localization accuracy, time to complete, situation awareness, workload.

Lessons Learned: Participants with "shared" mode found more items. Participants with "target" mode had best localization.

3.4 Multiple robot supervisory control [10]

Purpose: Evaluate metrics relating to the performance of a human-robot team with different numbers of robots.

Social robot or tool? Tool.

Test: Set of systems.

Participants: (12) No Robot – No Domain (Students).

Dependent Measures: Interaction time, interaction impact, neglect tolerance, wait time, number of tasks completed, switch times.

Lessons Learned: Developed metrics indicate the limits of the agents in the team and contain key performance parameters but fall short in the category of predictive power.

3.5 Search and rescue [6]

Purpose: To examine operator situation awareness and technical search team interaction using communication analysis.

Social robot or tool? Tool.

Participants: (5) Robot – Domain (USAR students, instructors).

Test: System.

Dependent Measures: Communications between the robot operators and other team members, situation awareness.

Lessons Learned: Operators spent more time gathering information about the state of the robot and the environment than moving the robot. Operators had difficulty integrating the robot's perspective into their own view of the search and rescue sight.

3.6 Search and rescue II [9]

Purpose: This report was not based on an experiment, but on a real un-staged rescue response to the attack on the World Trade Center in September 2001.

Social robot or tool? Tool. Also can be viewed as a social robot from the perspective of an essential teammate in the rescue response.

Participants: Robot – Domain (USAR professionals).

Test: System.

Dependent Measures: Human-robot ratios, objects identified, observe, how robots were used, drop time.

Lessons Learned: Most pressing needs to improve HRI are to reduce transport and operator human-robot ratios, create intelligent and assistive interfaces, and dedicate user studies to identify issues in the social niche of robotics applied to the USAR domain.

3.7 Interfaces for robot search tasks [33]

Purpose: To learn which interface design elements are most useful to USAR personnel while performing a search task.

Social robot or tool? Tool.

Participants: (8) No Robot – Domain (USAR professionals).

Test: System.

Dependent Measures: Area coverage, collisions, victims found.

Lessons Learned: Participants covered slightly more of the area with the INL system. Operators preferred the INL 3D map, but did not like the INL overlay of the map on top of the video.

3.8 Interfaces for robot search tasks II [33]

Purpose: To compare a new interface design with previous interface designs in a maze navigation task.

Social robot or tool? Tool.

Participants: (18) No Robot – No Domain (Students).

Test: Component.

Dependent Measures: Time to completion, collisions.

Lessons Learned: 3D map perspective yielded fewer collisions and faster time to complete.

3.9 Interactive robot in a nursing home [29]

Purpose: To study the effects of human-robot interaction between a robot and seniors in a nursing home.

Social robot or tool? Social. The authors are examining the relationship between the human and the robot.

Participants: No Robot – Domain (Elderly who have not interacted with a robot previously).

Test: System.

Dependent Measures: Level of social activity.

Lessons Learned: Participants were more active when the robot was on, than when it was off suggesting that the robotic aspect of the creature improved activity among participants.

3.10 Dance interaction in a classroom [30]

Purpose: To observe and evaluate the effects of two dance algorithms on a classroom of children and to experiment with different methods for evaluating the interaction developed between the children and the robot.

Social robot or tool? Social. The authors are examining how the class responds to the robot.

Participants: No Robot – Domain (Young children who have not previously had a robotic playmate).

Test: Algorithm, System.

Dependent Measures: "good child-robot interactions", number of children in the room, frequency of children re-entering the room, duration of time with robot.

Lessons Learned: Children spent longer each day in the room when there was a robot present rather than when only music was playing. As time progressed, children changed how they interacted with the robot based on their perceived weaknesses (e.g. easily falls down so handle softly).

4. METRIC CLASSES

From the previous section we see that there are numerous dependent measures used by a variety of experimenters to come up with ways of evaluating their human-robot solution. In fact, it could be said that there is no need to come up with new or more powerful metrics. Rather, perhaps our focus should be spent on classifying the metrics so we can gain a better understanding of what is being measured and how it relates to work done from other research institutions. In this section we will discuss two commonly used metric classes that address a) the evaluation of the task objective and b) the human involvement with the robot system). While these two classes of metrics account for most of the effort in a human-robot interaction system, there is also a third class of metrics that account for the reliability of a system. It seems that in many HRI reports, the reliability of the solution is rarely discussed, despite the affect of reliability on operator's trust in the system and even how the operator ends up using the system. Therefore, this section will also discuss the role of reliability metrics and how they apply to the domain of human-robot interactions.

4.1 Metrics evaluating task objective

The class of metrics that evaluate the performance of a system relative to the task objective is focused on determining how well the task, goal, or mission is achieved. With respect to the field of human-robot interaction, measures of performance in

this metric class can usually be achieved even for systems that do not have a robotic element. For example, in a hazardous material detection exercise, response personnel may be evaluated on how quickly they searched and mapped an area, how accurately they reported the location of hazardous material and how accurately they followed their procedures. The important thing is that these measures of performance could be achieved with or without robots as part of the tools available to the response personnel.

Some of the measures of performance that relate to the class of evaluating the task objective include: time to completion, time to accomplish a subtask, errors made during the task, safety of the task, correlation of task performance with mission objectives, number of items found, correct classification of sensory input, coverage of the environment, similarity to human-performance, or tradeoffs between human and robot to accomplish the task.

4.2 Metrics of human involvement

The class of metrics of human involvement includes the measurement of human activities with respect to the operation of the robot. This class of metrics is primarily used to improve our understanding about the relationships between user input and task or objective performance as discussed in the previous section. In some cases, performance is measured, and then reasons are sought by looking at the interaction data. In other cases, theories of interaction are explored and hypothesized then experiments designed to evaluate the effect of the theory on performance.

Much of the literature in HRI has been focused on metrics in this area, for example, Crandall and Cummings have identified two sub-classes of human involvement that include interaction efficiency (how effectively the operator interacts with the robot) and neglect efficiency (how well the robot maintains levels of performance when neglected by the operator) [10]. In fact their separation of metrics into two classes is particularly beneficial to most HRI activities because it separates metrics relating to the role of the operator from those relating to the role of the robot.

Measures of performance that relate to interaction efficiency might include usability, interaction time joystick bandwidth, number of teloperation commands, NASA-TLX workload, Behavioral entropy workload, number of teleoperations, number of backups, mouse movement, joystick movement, keyboard presses, proximal interactions, or team interactions.

Measures of performance that relate to the class of neglect tolerance might include, autonomy level of the robot, interaction scheme, percentage of time in autonomy modes, reasons for changing autonomy modes, span of control and fan-out.

Another common subclass of metrics related to human involvement and widely discussed in relation to HRI is that of situation awareness. The evaluation of situation awareness as applied to robotics most likely originated from Endsley's work with aviation pilots who identified the "edge" that some pilots had in aerial combat and other aviation tasks as an improvement of their "situational awareness." A commonly cited definition of situation awareness is "The perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [13]. Many researchers have discussed situation awareness as a means to better understand how well and how accurately the operator understands the robot's relationship with the environment and the operator [6, 26, 27].

4.3 Metrics of system reliability

One class of metrics that seems to be missing from many HRI reports and discussions is that of system reliability. By this we mean measures that indicate how consistently the human-robot system works. Moreover, this class of metrics can be used to clarify the nuances, both good and bad, of a particular system.

Some may discount the value of this class of metrics to the HRI field and claim that it belongs more in the engineering field of robotics as it relates more to the actual engineering (mechanical, electrical, or computer) of the robot. However, we claim that it belongs in HRI evaluations because there must be a means to classify the nuances of human responses to robot systems that sometimes fail, even if the failure is related to the robustness of the hardware, software, or implementation of the algorithms. While this class of metrics might not be interesting to many researchers in HRI, it is none the less very important to the end user. As an example, consider the following anecdotal story.

In a recent experiment involving military soldiers, we were having some trouble with our iRobot PackBot where the video was being broadcast as a blank red image. As developers we had never seen this and therefore had a hard time identifying the cause. This red image caused significant delay in our experiments for a few days and we were left to converse with the soldiers. We asked them if they had ever seen anything like it before. They said, they had and their solution was to shut down the system and restart it three or four times, then it would usually go away. We later discovered that the cause of the red video was a hairline fracture in the video capture board and when the robot was cold, the electronic connection was not established however, when the robot warmed up through use, the slight expansion of the connections on the board caused the video to work. If the robot used by the soldiers had the same problem, then restarting the system would be the wrong approach to use because it would only delay the "warming-up" of the system.

In how many human-robot experiments do similar stories unfold? While this can often be comical to the people participating in the experiment and a nuisance to the experimenter, such experiences can have a lasting effect on potential end-users as we work to bring robotics to domains such as search and rescue and the military wherein these experiences may cause users to view the robot as shoddy equipment that has some value, but that they wouldn't be able to depend on. In some ways, it is like a flashlight or rifle that worked 99 out of 100 times, but 1% of the time, would not turn on or not fire and required the operator to remove the batteries or the bullets and re-install them—manageable yes, but certainly not preferred. These issues certainly affect how the human uses and interacts with the robot.

The issue is not that entire papers need to be written on the reliability aspects of the HRI systems, rather, it would be beneficial for both readers and researchers to have some understanding regarding the reliability or robustness of the system being used. For example, did the experiment have to be halted for any reason related to the technology? How reliable is the software, hardware, communications, and/or algorithms? How much power does the robot system use? Did the robot perform unsafe actions? What about communication bandwidth or computational resource requirements? Information regarding these questions could be short and to the point but would provide valuable insights into the reliability of the human-robot solution.

Another important aspect with respect to reliability is to consider how terminology is defined and used. For example, it seems that many institutions have their own description of levels of dynamic autonomy. For example, shared control might involve continuous joystick input [11] or infrequent joystick input [5], or haptic input [18].

Related to the issues of defining modes of autonomy is the question of autonomy mode performance. If there is a "guarded-motion" behavior on the robots, then how are collisions with obstacles recorded? Are they recorded as errors in operator situation awareness or in the reliability of the guarded motion? Moreover, why do some autonomy systems make a task faster while others make a task slower? Answers to these questions may be related to the reliability and craftsmanship of the autonomy modes themselves rather than the general principles of each autonomy mode.

In order to fully understand the contributions of a human-robot interaction experiment, it is beneficial to understand some of the core questions relating to the reliability of the system. This can help others verify the results at their own laboratory and make meaningful comparisons between experiments performed by different researchers.

5. CONCLUSIONS

This paper presented a simplified framework for characterizing experiments in human-robot interactions with the intent of facilitating comparisons between experiments from different research institutions and understanding the contributions from new research. The framework categorizes experiments by asking questions about the purpose of the robot, what is being evaluated, and the types of participants in the experiment. Experiments with similar answers to these questions should be grouped together and meaningful comparisons between the experiments should exist. The measures of performance related with experiments are also grouped into three broad metric classes that include a) evaluations of the task objective, b) human involvement with the robot system, and c) system reliability.

While system reliability is often overlooked in reports on human-robot interaction, it is essential to understand the reliability of the human-robot system because the reliability affects how end users will actually interact with the system as it becomes integrated into their daily routine.

The research in this paper focused primarily on applying the framework to robots that are used as tools in critical environments. Future work needs to show how social robots fit into this same framework.

6. REFERENCES

[1] Ambrose, R., Aldridge, H., Askew, R., Burridge, R., Bluethmann, W., Diftler, M. Lovchik, C. Magruder, D. and Rehnmark, F. 2000. Robonaut: NASA's space humanoid. IEEE Intelligent Systems. 15, 5, 57-63.

[2] Breazeal, C. 2002. Designing sociable robots. The MIT Press, Cambridge, MA.

[3] Breazeal C. 2003. Towards sociable robots. T. Fong (ed), Robotics and Autonomous Systems, 42, 3-4, 167-175.

[4] Bruemmer, D. J., Nielsen, C. W., Gertman, D. I. 2008. How training and experience affect the benefits of autonomy in a dirty-bomb experiment. Accepted for publication at the 3rd ACM/IEEE International Conference on Human-Robot Interaction (Amsterdam, The Netherlands, March 12-15, 2008).

[5] Bruemmer, D. J., Few, D. A., Boring, R. L., Marble, J. L., Walton, M. C., and Nielsen C. W. 2005. Shared understanding for collaborative control. IEEE Transactions on Systems, Man, and Cybernetics, Part A. Systems and Humans (July 2005), 35, 4, 505-512.

[6] Burke, J. L., Murphy, R. R., Coovert, M. D., and Riddle, D. 2004. Moonlight in Miami: Field Study of Human-Robot Interaction in the Context of an Urban Search and Rescue Disaster Response Training Exercise. Human-Computer Interaction, 19, 1&2, 85-116.

[7] Burke, J. L., Murphy, R. R., Rogers, E. Lumelsky, V. J., and Scholtz, J. 2004. Final report for the DARPA/NSF interdisciplinary study on human-robot interaction. IEEE Transactions on Systems, Man, and Cybernetics, Part C (May 2004), 34, 2, 103–112.

[8] Byers, Z., Dixon, M., Goodier, K. Grimm, C. M. and Smart, W. D. 2003. An autonomous robot photographer. In Proceedings of the International Conference on Robots and Systems (IROS 2003, Las Vegas, NV, October, 2003), 2636-2641.

[9] Casper, J. and Murphy, R. R. 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 33, 3, 367-385.

[10] Crandall, J. W. and Cummings, M. L. 2007. Developing performance metrics for the supervisory control of multiple robots. Proceedings of the 2nd ACM/IEEE International conference on Human-Robot Interaction (HRI 2007, Arlington, VA).

[11] Crandall, J. W. and Goodrich, M. A. 2002. Characterizing efficiency of human-robot interaction: A case study of shared-control teleoperation. In proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems.

[12] Duffy, B. 2003. Anthropomorphism and the social robot. Robotics and Autonomous Systems (March 2003), 42, 3-4, 177-190.

[13] Endsley, M. R. 1988. Design and evaluation for situation awareness enhancement. In proceedings of the Human Factors Society 32nd Annual Meeting, 97-101.

[14] Hill, S. G. and Bodt, B. 2007. A field experiment of autonomous mobility: operator workload for one and two robots. Proceedings of the 2nd ACM/IEEE International conference on Human-Robot Interaction (HRI 2008, Arlington, VA).

[15] Jacobson, M. J. 2000. Problem solving about complex systems: Differences between experts and novices. In B. Fishman & S. O'Connor-Divelbiss (Eds.), Fourth International Conference of the Learning Sciences, 14-21. Mahwah, NJ: Erlbaum.

[16] Jones H., Rock, S., Burns, D., and Morris, S. 2002. Autonomous robots in SWAT applications: Research, design, and operatiosn challenges. In Proceedings of the symposium for the Association of Unmanned Vehicle Systems Interantional (AUVSI, Orlando, FL, 2002).

[17] Kang, K. I., Freedman, S., and Matari´c, M. J. A hands-off physical therapy assistance robot for cardiac patients. In

proceedings of the IEEE International Conference on Rehabilitation Robotics (ICORR, Chicago, IL, June 2005).

[18] Lacey, G. and MacNamara S. 2000. Context-aware shared control of a robot mobility aid for the elderly blind. The International Journal of Robotics Research. 19, 11, 1054-1065.

[19] Lundberg, C., Christensen, H., and Reinhold, R., Long-term study of a portable field robot in urban terrain: Field reports. Journal of Field Robotics (August 2007), 24, 8-9, 625-650.

[20] Murphy. R. R. 2004. Human-robot interaction in rescue robotics. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 34, 2, 138-153.

[21] Nielsen, C. W. and Bruemmer, D. J. 2007. Hiding the system from the user: moving from complex mental models to elegant metaphors. In proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication (Ro-Man August, 2007, Jeju Island, Korea).

[22] Nielsen, C. W., Goodrich M. A., and Ricks, B. 2007. Ecological interfaces for improving mobile robot teleoperation. IEEE Transactions on Robotics (October 2007), 23, 5, 927-941.

[23] Olsen, D. R. and Wood, B. 2004. Fan-out: Measuring human control of multiple robots. Proceedings of CHI 2004, ACM.

[24] Pedahazur, E. J. and Pedhazur Schmelkin, L. 1991. Measurement, design, and analysis: An integrated approach. Hillsdale, N.J. Lawrence Erlbaum Associates.

[25] Reeves, B. and Nass, C. 1996. The Media Equation. CSLI Publications, Stanford, CA.

[26] Scholtz, J., Young, J., Drury, J. L., and Yanco, H. A. Evaluation of human-robot interaction awareness in search and rescue. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA April 2004, New Orleans).

[27] Sellner, B. P., Hiatt, L. M., Simmons, R., and Singh, S. 2006. Attaining situation awareness for sliding autonomy. In proceedings of the 1st annual conference on Human-robot interaction (HRI, March 2006, Salt Lake City, UT), 80-87.

[28] Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J. Scultz, A., and Goodrich, M. 2006. Common metrics for human-robot interaction. Proceedings of the 2nd ACM/IEEE International conference on Human-Robot Interaction (HRI 2007, Arlington, VA).

[29] Taggart, W., Turkle, S., and Kidd, C. D. An interactive robot in a nursing home: Preliminary remarks. In Proceedings of the CogSci – 2005 Workshop: Toward Social Mechanisms of Android Science, Stresa, Italy, July 2005.

[30] Tanaka, F., Movellan, J. R., Fortenberry, B., and Aisaka, K. Daily HRI evaluation at a classroom environment: Reports from dance interaction experiments. Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI, March 2006, Salt Lake City, UT) 3-9.

[31] Thrun S., Beetz, M., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Hähnel, D., Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. 2000. Probabilistic algorithms and the interactive museum tour-guide robot MINERVA. Journal of robotics research, 19, 11, 972-999.

[32] Yanco, H. A. and Drury, J. L. 2004. Classifying human-robot interaction: An updated taxonomy. Proceedings of the IEEE Conference on Systems, Man and Cybernetics(October 2004, The Hague, The Netherlands).

[33] Yanco, H. A., Keyes, B., Drury, J. L. Nielsen, C. W., Few, D. A., Bruemmer, D. J. 2007. Evolving interface design for robot search tasks. Journal of Field Robotics (August/September 2007), 24, 8/9, 779-799.

Measuring the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots.

Christoph Bartneck
Department of Industrial Design
Eindhoven University of Technology
Den Dolech 2, 5600MB Eindhoven
The Netherlands
Phone +31 40 247 5175
c.bartneck@tue.nl

Dana Kulic
Nakamura & Yamane Lab
Department of Mechano-Informatics
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-8656, Japan
dana@ynl.t.u-tokyo.ac.jp

Elizabeth Croft
Department of Mechanical Engineering
University of British Columbia
6250 Applied Science Lane
Room 2054, Vancouver
Canada V6T 1Z4
ecroft@mech.ubc.ca

ABSTRACT
This study emphasizes the need for standardized measurement tools for human robot interaction (HRI). If we are to make progress in this field then we must be able to compare the results from different studies. A literature review has been performed on the measurements of five key concepts in HRI: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. The results have been distilled into five consistent questionnaires using semantic differential scales. We report reliability and validity indicators based on several empirical studies that used these questionnaires. It is our hope that these questionnaires can be used by robot developers to monitor their progress. Psychologists are invited to further develop the questionnaires by adding new concepts, and to conduct further validations where it appears necessary.

Categories and Subject Descriptors
H.5.1 [**Multimedia Information Systems**]: Evaluation/methodology

General Terms
Measurement, Human Factors, Standardization

Keywords
Human factors, robot, perception, measurement.

1. INTRODUCTION
The success of service robots and, in particular, of entertainment robots cannot be assessed only by performance criteria typically found for industrial robots. The number of processed pieces and their accordance with quality standards are not necessarily the prime objectives for an entertainment robot such as Aibo (Sony, 1999), or a communication platform such as iCat (Breemen, Yan, & Meerbeek, 2005). The performance criteria of service robots lie within the satisfaction of their users. Therefore, it is necessary to measure the users' perception of service robots, since these can not be measured within the robots themselves.

Measuring human perception and cognition has its own pitfalls, and psychologists have developed extensive methodologies and statistical tests to objectify the acquired data. Most engineers who develop robots are often unaware of this large body of knowledge, and sometimes run naïve experiments in order to verify their designs. But the same naivety can also be expected of psychologists when confronted with the task of building a robot. Human-Robot Interaction (HRI) is a multidisciplinary field, but it can not be expected that everyone masters all skills equally well. We do not intend to investigate the structure of the HRI community and the problems it is facing in the cooperation of its members. The interested reader may consult Bartneck & Rauterberg (Bartneck & Rauterberg, 2007) who reflected on the structure of the Human-Computer Interaction community This may also apply to the HRI community. This study is intended for the technical developers of interactive robots who want to evaluate their creations without having to take a degree in experimental psychology. However, it is advisable to at least consult with a psychologist over the overall methodology of the experiment.

A typical pitfall in the measurement of psychological concepts is to break them down into smaller, presumably better-known, components. This is common practice, and we do not intend to single out a particular author, but we still feel the need to present an example. Kiesler and Goetz (2002) divided the concept of anthropomorphism into the sub components sociability, intellect, and personality. They measured each concept with the help of a questionnaire. This breaking down into sub components makes sense if the relationship and relative importance of the sub components are known and can therefore be calculated back into the original concept. Otherwise, a presumably vague concept is simply replaced by series of just as vague concepts. There is no reason to believe that it would be easier for the users of robots to evaluate their sociability rather than their anthropomorphism. Caution is therefore necessary so as not to over-decompose concepts. Still, it is good practice to at least decompose the concept under investigation into several items[1] so as to have richer and more reliable data as was suggested by Fink, volume 8, p. 20 (2003).

A much more reliable and possibly objective method for measuring the users' perception and cognition is to observe their behavior. If, for example, the intention of a certain robot is to play a game with the user, then the fun experienced can be deduced from the time the user spends playing it. The longer the user plays, the more fun it is. However, not all internal states of a user manifest themselves in observable behavior. From a practical point of view it can also be very laborious to score the users' behaviors on the basis of video recordings.

Physiological measurements form a second group of measurement tools. Skin conductivity, heart rate, and heart variance are three popular measurements that provide a good indication of the user's arousal in real time. The measurement can be taken during the interaction with the robot. Unfortunately, these measurements can not distinguish the arousal that stems from anger from that which may originate from joy. To gain better insight into the user's state, these measurements can be complemented by other physiological measurements, such as the recognition of facial expression. In combination, they can provide real time data, but the effort of

[1] In the social sciences the term "item" refers to a single question or response.

setting up and maintaining the equipment and software should not be underestimated.

A third measurement technique is questionnaires, which are often used to measure the users' attitudes. While this method is rather quick to conduct, its conceptual pitfalls are often underestimated. One of its prime limitations is, of course, that the questionnaire can be administered only after the actual experience. Subjects have to reflect on their experience afterwards, which might bias their response. They could, for example, adapt their response to the socially acceptable response.

The development of a validated questionnaire involves a considerable amount of work, and extensive guidelines are available to help with the process (Dawis, 1987; Fink, 2003). Development will typically begin with a large number of items, which are intended to cover the different facets of the theoretical construct to be measured; next, empirical data is collected from a sample of the population to which the measurement is to be applied. After appropriate analysis of this data, a subset of the original list of items is then selected and becomes the actual multi-indicator measurement. This measurement will then be formally assessed with regard to its reliability, dimensionality, and validity.

Due to their naivety and the amount of work necessary to create a validated questionnaire, developers of robots have a tendency to quickly cook up their own questionnaires. This conduct results in two main problems. Firstly, the validity and reliability of these questionnaires has often not been evaluated. An engineer is unlikely to trust a voltmeter developed by a psychologist unless its proper function has been shown. In the same manner, psychologists will have little trust in the results from a questionnaire developed by an engineer unless information about its validity and reliability is available. Secondly, the absence of standard questionnaires makes it difficult to compare the results from different researchers. If we are to make progress in the field of human-robot interaction then we shall have to develop standardized measurement tools similar to the ITC-SOPI questionnaire that was developed to measure presence (Lessiter, Freeman, Keogh, & Davidoff, 2001).

This study attempts to make a start in the development of standardized measurement tools for human-robot interaction by first presenting a literature review on existing questionnaires, and then presenting empirical studies that give an indication of the validity and reliability of these new questionnaires. This study will take the often-used concepts of anthropomorphism, animacy, likeability, and perceived intelligence and perceived safety as starting points to propose a consistent set of five questionnaires for these concepts.

We can not offer an exhaustive framework for the perception of robots similar to the frameworks that have already been developed for social robots (Bartneck & Forlizzi, 2004; Fong, Nourbakhsh, & Dautenhahn, 2003) that would justify the selection of these five concepts. We can only hint at the fact that the concepts proposed have been necessary for our own research and that they are likely to have relationships with each other. A highly anthropomorphic and intelligent robot is likely to be perceived to be more animate and possibly also more likeable. The verification of such a model does require appropriate measurement instruments. The discussion of whether it is good practice to first develop a theory and then the observation method or vice versa has not reached a conclusion (Chalmers, 1999), but every journey begins with a first step. The proposed set of questionnaires can later be extended to cover other relevant concepts, and their relationships can be further explored. The emphasis is on presenting questionnaires that can be used directly in the development of interactive robots. Many robots are being built right now, and the engineers cannot wait for a mature model to emerge. We even seriously consider the position that such a framework can be created only once we have the robots and measurement tools in place.

Unfortunately, the literature review revealed questionnaires that used different types of items, namely Likert-scales (Likert, 1932) and semantic differential scales (Osgood, Suci, & Tannenbaum, 1957). If more than one questionnaire is to be used for the evaluation of a certain robot, it is beneficial if the questionnaires use the same type of items. This consistency makes it easy for the participants to learn the method and thereby avoids errors in their responses. It was therefore decided to transfer Likert type scales to semantic differential scales. We shall now discuss briefly the differences between these two types of items.

In semantic differential scales the respondent is asked to indicate his or her position on a scale between two bipolar words, the anchors (see Figure 1, top). In Likert scales (see Figure 1, bottom), subjects are asked to respond to a stem, often in the form of a statement, such as "I like ice cream". The scale is frequently anchored with choices of "agree" - "disagree" or "like" - "dislike".

Figure 1. Example of a semantic differential scale (top) and likert scale (bottom). The participant would be asked to rate the stimulus on this scale by circling one of the numbers.

Both are rating scales, and provided that response distributions are not forced, semantic differential data can be treated just as any other rating data (Dawis, 1987). The statistical analysis is identical. However, a semantic differential format may effectively reduce acquiescence bias without lowering psychometric quality (Friborg, Martinussen, & Rosenvinge, 2006). A common objection to Osgood's semantic differential method is that it appears to assume that the adjectives chosen as anchors mean the same to everyone. Thus, the method becomes self-contradictory; it starts from the presumption that different people interpret the same word differently, but has to rely on the assumption that this is not true for the anchors. However, this study proposes to use the semantic differential scales to evaluate not the meaning of words, but the attitude towards robots. Powers and Kiesler (2006) report a negative correlation (-.23) between Humanlikeness and Machinelikeness, which strengthens our view that semantic differentials are a useful tool for measuring the users' perception of robots, while we remain aware of the fact that every method has its limitations.

Some information on the validity and reliability of the questionnaires is already available from the original studies on which they are based. However, the transformation from Likert scales to semantic differential scales may compromise these indicators to a certain degree. We shall compensate this possible loss by reporting on complementary empirical studies later in the text. First, we would like to discuss the different types of validity and reliability.

Fink in Volume 8, pp 5-44, (Fink, 2003) discusses several forms of reliability and validity. Among the scientific forms of

validity we find content validity, criterion validity, and construct validity. The latter, which determines the degree to which the instrument works in comparison with others, can only be assessed after years of experience with a questionnaire, and construct validity is often not calculated as a quantifiable statistic. Given the short history of research in HRI it would appear difficult to achieve construct validity. The same holds true for criterion validity. There is a scarcity of validated questionnaires with which our proposed questionnaires can be compared. We can make an argument for content validity since experts in the field carried out the original studies, and measurements of the validity and reliability have even been published from time to time. The researchers involved in the transformation of the proposed questionnaires were also in close contact with relevant experts in the field with regard to the questionnaires. The proposed questionnaires can therefore be considered to have content validity.

It is easier to evaluate the reliability of the questionnaire, and Fink describes three forms: test-retest reliability, alternate form reliability, and internal consistency reliability. The latter is a measurement for how well the different items measure the same concept, and it is of particular importance to the questionnaires proposed because they are designed to be homogenous in content. Internal consistency involves the calculation of a statistic known as Cronbach's Alpha. It measures the internal consistency reliability among a group of items that are combined to form a single scale. It reflects the homogeneity of the scale. Given the choice of homogeneous semantic differential scales, alternate form reliability appears difficult to achieve. The items cannot simply be negated and asked again because semantic differential scales already include dichotomous pairs of adjectives. Test-retest reliability can even be tested within the same experiment by splitting the participants randomly into two groups. This procedure requires a sufficiently large number of participants and unfortunately none of the studies that we have access to had enough participants to allow for a meaningful test-retest analysis. For both, test-retest reliability and internal consistency reliability, Nunnally (1978) recommends a minimum value of 0.7. We would now like to discuss the five concepts of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety in more detail, and describe a questionnaire for each of them.

2. ANTHROPOMORPHISM

Anthropomorphism refers to the attribution of a human form, human characteristics, or human behavior to nonhuman things such as robots, computers, and animals. Hiroshi Ishiguro, for example, develops androids that, for a short period, are indistinguishable from human beings (Ishiguro, 2005). His highly anthropomorphic androids struggle with the so-called 'uncanny valley', a theory that states that as a robot is made more humanlike in its appearance and movements, the emotional response from a human being to the robot becomes increasingly positive and empathic, until a point is reached beyond which the response quickly becomes that of intense repulsion. However, as the appearance and movements continue to become less distinguishable from those of a human being, the emotional response becomes positive once more and approaches human-human empathy levels.

Even if it is not the intention of the design of a certain robot to be as humanlike as possible, it still remains important to match the appearance of the robot with its abilities. A too anthropomorphic appearance can evoke expectations that the robot might not be able to fulfill. If, for example, the robot has a human-shaped face then the naïve user will expect that the robot is able to listen and to talk. To prevent disappointment it is necessary for all developers to pay close attention to the anthropomorphism level of their robots.

An interesting behavioral measurement for anthropomorphism has been presented by Minato et al. (2005). They attempted to analyze differences in where the participants were looking when they looked at either a human or an android. The hypothesis is that people look differently at humans compared to robots. They have not been able to produce reliable conclusions yet, but their approach could turn out to be very useful, assuming that they can overcome the technical difficulties.

MacDorman (2006) presents an example of a naïve questionnaire. A single question is asked to assess the human-likeness of what is being viewed (9-point semantic differential, mechanical versus humanlike). It is good practice in the social sciences to ask multiple questions about the same concept in order to be able to check the participants' consistency and the questionnaire's reliability. Powers and Kiesler (2006), in comparison, used six items and are able to report a Cronbach's Alpha of 0.85. Their questionnaire therefore appears to be more suitable. It was necessary to transform the items used by Powers and Kiesler into semantic differentials: Fake / Natural, Machinelike / Humanlike, Unconscious / Conscious, Artificial / Lifelike, and Moving rigidly / Moving elegantly.

Two studies are available in which this new anthropomorphism questionnaire was used. The first one reports a Cronbach's Alpha of 0.878 (Bartneck, Kanda, Ishiguro, & Hagita, 2007) and we would like to report the Cronbach's Alphas for the second study (Bartneck, Kanda, Ishiguro, & Hagita, 2008) in this paper. The study consisted of three within conditions for which the Cronbach's Alphas must be reported separately. We can report a Cronbach's Alpha of 0.929 for the human condition, 0.923 for the android condition and 0.856 for the masked android condition. The alpha values are well above 0.7, so we can conclude that the anthropomorphism questionnaire has sufficient internal consistency reliability.

3. ANIMACY

The goal of many robotics researchers is to make their robots lifelike. Computer games, such as The Sims, Creatures, or Nintendo Dogs show that lifelike creatures can deeply involve users emotionally. This involvement can then be used to influence users (Fogg, 2003). Since Heider and Simmel (1944), a considerable amount of research has been devoted to the perceived animacy and "intentions" of geometric shapes on computer screens. Scholl and Tremoulet (2000) offer a good summary of the research field, but, on examining the list of references, it becomes apparent that only two of the 79 references deal directly with animacy. Most of the reviewed work focuses on causality and intention. This may indicate that the measurement of animacy is difficult. Tremoulet and Feldman (2000) only asked their participants to evaluate the animacy of 'particles' under a microscope on a single scale (7-point Likert scale, 1=definitely not alive, 7=definitely alive). It is questionable how much sense it makes to ask participants about the animacy of particles. By definition they cannot be alive since particles tend to be even smaller than the simplest organisms.

Asking about the perceived animacy of a certain stimulus makes sense only if there is a possibility for it to be alive. Robots can show physical behavior, reactions to stimuli, and even language skills. These are typically attributed only to animals, and hence it can be argued that it makes sense to ask participants about their perception of the animacy of robots.

McAleer, et al. (2004) claim to have analyzed the perceived animacy of modern dancers and their abstractions on a computer screen, but only qualitative data of the perceived arousal is presented. Animacy was measured with free responses. They looked for terms and statements that indicated that subjects had attributed human movements and characteristics to the shapes. These were terms such as "touched", "chased", and "followed", and emotions such as "happy" or "angry". Other guides to animacy were when the shapes were generally being described in active roles, as opposed to being controlled in a passive role. However, they do not present any quantitative data for their analysis.

A better approach has been presented by Lee, Kwan Min, Park, Namkee & Song, Hayeon (2005). With their four items (10-point Likert scale; lifelike, machine-like, interactive, responsive) they have been able to achieve a Cronbach's Alpha of 0.76. For the questionnaires in this study, their items have been transformed into semantic differentials: Dead / Alive, Stagnant / Lively, Mechanical / Organic, Artificial / Lifelike, Inert / Interactive, Apathetic / Responsive. One study used this new questionnaire (Bartneck, Kanda, Mubin, & Mahmud, 2007) and reported a Cronbach's Alpha of 0.702, which is sufficiently high for us to conclude that the new animacy questionnaire has sufficient internal consistency reliability.

4. LIKEABILITY

It has been reported that the way in which people form positive impressions of others is to some degree dependent on the visual and vocal behavior of the targets (Clark & Rutter, 1985), and that positive first impressions (e.g., likeability) of a person often lead to more positive evaluations of that person (Robbins & DeNisi, 1994). Interviewers report knowing within 1 to 2 minutes whether a potential job applicant is a winner, and people report knowing within the first 30 seconds the likelihood that a blind date will be a success (Berg & Piner, 1990). There is a growing body of research indicating that people often make important judgments within seconds of meeting a person, sometimes remaining quite unaware of both the obvious and subtle cues that may be influencing their judgments. Since computers, and thereby robots in particular, are to some degree treated as social actors (Nass & Reeves, 1996), it can be assumed that people are able to judge robots just as.

Jennifer Monathan (1998) complemented her "liking" question with 5-point semantic differential scales: nice / awful, friendly / unfriendly, kind / unkind, and pleasant / unpleasant, because these judgments tend to demonstrate considerable variance in common with "liking" judgments (Burgoon & Hale, 1987). Monahan later eliminated the kind-unkind and pleasant-unpleasant items in her own analysis since they did not load sufficiently in a factor analysis that also included items from three other factors. The Cronbach's Alpha of 0.68 therefore relates only to this reduced scale. Her experimental focus is different from the intended use of her questionnaire in the field of HRI. She also included concepts of physical attraction, conversational skills, and other orientations, which might not be of prime relevance to HRI. In particular, physical attraction might be unsuitable for robots. No reports on successful human-robot reproduction are available yet and hopefully never will be. We decided to only include the five items, since it is always possible to exclude items in cases where they would not contribute to the reliability and validity of the questionnaire.

Two studies used this new likeability questionnaire. The first reports a Cronbach's Alpha of 0.865 (Bartneck, Kanda, Ishiguro, & Hagita, 2007), and we report the Cronbach's Alpha for the second (Bartneck, Kanda, Ishiguro, & Hagita, 2008) in this paper. The study consisted of three "within" conditions for which the Cronbach's Alpha must be reported separately. Without going into too much detail of the study, we can report a Cronbach's Alpha of 0.923 for the human condition, 0.878 for the android condition, and 0.842 for the masked android condition. The alpha values are well above 0.7, and hence we can conclude that the likeability questionnaire has sufficient internal consistency reliability.

5. PERCEIVED INTELLIGENCE

Interactive robots face a tremendous challenge in acting intelligently. The reasons can be traced back to the field of artificial intelligence (AI). The robots' behaviors are based on methods and knowledge that were developed by AI. Many of the past promises of AI have not been fulfilled, and AI has been criticized extensively (Dreyfus & Dreyfus, 1992; Dreyfus, Dreyfus, & Athanasiou, 1986; Searle, 1980; Weizenbaum, 1976).

One of the main problems that AI is struggling with is the difficulty of formalizing human behavior, for example, in expert systems. Computers require this formalization to generate intelligent and human-like behavior. And as long as the field of AI has not made considerable progress on these issues, robot intelligence will remain at a very limited level. So far, we have been using many Wizard-Of-Oz methods to fake intelligent robotic behavior, but this is possible only in the confines of the research environment. Once the robots are deployed in the complex world of everyday users, their limitations will become apparent. Moreover, when the users are interacting with the robot for years rather than minutes, they will become aware of the limited abilities of most robots.

Evasion strategies have also been utilized. The robot would show more or less random behavior while interacting with the user, and the user in turn sees patterns in this behavior which he/she interprets as intelligence. Such a strategy will not lead to a solution of the problem, and its success is limited to short interactions. Given sufficient time the user will give up his/her hypothesized patterns of the robot's intelligent behavior and become bored with its limited random vocabulary of behaviors. In the end, the perceived intelligence of a robot will depend on its competence (Koda, 1996). To monitor the progress being made in robotic intelligence it is important to have a good measurement tool.

Warner and Sugarman (1996) developed an intellectual evaluation scale that consists of five seven-point semantic differential items: Incompetent / Competent, Ignorant / Knowledgeable, Irresponsible / Responsible, Unintelligent / Intelligent, Foolish / Sensible. Parise et al. (Parise, Kiesler, Sproull , & Waters 1996) excluded one question from this scale, and reported a Cronbach's Alpha of 0.92. The questionnaire was again used by Kiesler, Sproull and Waters (Kiesler, Sproull, & Waters, 1996), but no alpha was reported. Three other studies used the perceived intelligence questionnaire, and reported Cronbach's Alpha values of 0.75 (Bartneck, Kanda, Ishiguro, & Hagita, 2008), 0.769 (Bartneck, Verbunt, Mubin, & Mahmud, 2007), and 0.763 (Bartneck, Kanda, Mubin, & Mahmud, 2007). These values are above the suggested 0.7 threshold, and hence the perceived intelligence questionnaire can be considered to have satisfactory internal consistency reliability.

6. PERCEIVED SAFETY

Perceived safety describes the user's perception of the level of danger when interacting with a robot, and the user's level of comfort during the interaction. Achieving a positive perception of safety is a key requirement if robots are to be accepted as partners and co-workers in human environments. Perceived

safety and user comfort have rarely been measured directly. Instead, indirect measures have been used - the measurement of the affective state of the user through the use of physiological sensors (Kulic & Croft, 2005; Rani, Sarkar, Smith, & Kirby, 2004; Rani, Sims, Brackin, & Sarkar, 2002), questionnaires (Inoue, Nonaka, Ujiie, Takubo, & Arai, 2005; Kulic & Croft, 2005; Wada, Shibata, Saito, & Tanie, 2004), and direct input devices (Koay, Walters, & Dautenhahn, 2005). That is, instead of asking subjects to evaluate the robot, researchers frequently use affective state estimation or questionnaires asking how the subject feels in order to measure the perceived safety and comfort level indirectly.

For example, Sarkar proposes the use of multiple physiological signals to estimate affective state, and to use this estimate to modify robotic actions to make the user more comfortable (Sarkar, 2002). Rani et al. (2004; 2002) use heart-rate analysis and multiple physiological signals to estimate human stress levels. In Rani et al. (2004), an autonomous mobile robot monitors the stress level of the user, and if the level exceeds a certain value, the robot returns the user in a simulated rescue attempt. However, in their study, the robot does not interact directly with the human; instead, pre-recorded physiological information is used to allow the robot to assess the human's condition.

Koay et al. (2005) describe an early study where human reaction to robot motions was measured online. In this study, 28 subjects interacted with a robot in a simulated living room environment. The robot motion was controlled by the experimenters in a "Wizard of Oz" setup. The subjects were asked to indicate their level of comfort with the robot by means of a handheld device. The device consisted of a single slider control to indicate comfort level, and a radio signal data link. Data from only 7 subjects was considered reliable, and was included in subsequent analysis. Analysis of the device data with the video of the experiment found that subjects indicated discomfort when the robot was blocking their path, the robot was moving behind them, or the robot was on a collision course with them.

Nonaka et al (2004) describe a set of experiments where human response to pick-and-place motions of a virtual humanoid robot is evaluated. In their experiment, a virtual reality display is used to depict the robot. Human response is measured through heart rate measurements and subjective responses. A 6-level scale is used from 1 = "never" to 6 = "very much", for the categories of "surprise", "fear", "disgust", and "unpleasantness". No relationship was found between the heart rate and robot motion, but a correlation was reported between the robot velocity and the subject's rating of "fear" and "surprise". In a subsequent study (Inoue, Nonaka, Ujiie, Takubo, & Arai, 2005), a physical mobile manipulator was used to validate the results obtained with the virtual robot. In this case, subjects are asked to rate their responses on the following (5-point) direction levels: "secure – anxious", "restless – calm", "comfortable – unpleasant", "unapproachable – accessible", "favorable – unfavorable", "tense – relaxed", "unfriendly – friendly", "interesting – tedious", and "unreliable – reliable". They are also asked to rate their level of "intimidated" and "surprised" on a 5 –point Likert scale. The study finds that similar results are obtained regardless of whether a physical or a virtual robot is used. Unfortunately, no information about the reliability or validity of their scales is available. There is a very large number of different questions that can be asked on the topic of safety and comfort in response to physical robot motion. This underlines the need for a careful and studied set of baseline questions for eliciting comparable results from research efforts, especially in concert with physiological measurement tools. It becomes apparent that two approaches can be taken to assess the perceived safety. On the one hand the users can be asked to evaluate their impression of the robot, and on the other hand they can be asked to assess their own affective state. It is assumed that if the robot is perceived to be dangerous then the user affective state would be tense.

Kulic and Croft (2005) combined a questionnaire with physiological sensors to estimate the user's level of anxiety and surprise during sample interactions with an industrial robot. They ask the user to rate their level of anxiety, surprise, and calmness during each sample robot motion. A 5 point Likert scale is used. The Cronbach's Alpha for the affective state portion of the questionnaire is 0.91. In addition, the subject is asked to rate their level of attention during the robot motion, to ensure that the elicited affective state was caused by the robot rather than by some other internal or external distraction. In this work, they show that motion planning can be used to reduce the perceived anxiety and surprise felt by subjects during high speed movements. This and later work (Kulic & Croft, 2006) by the same authors showed a strong statistical correlation between the affective state reported by the subjects and their physiological responses. The scales they produced can be transformed to the following semantic differential scales: Anxious / Relaxed, Agitated / Calm, Quiescent / Surprised. This questionnaire focuses on the affective state of the user. To our knowledge, no suitable questionnaire for rating the safety of a robot is available.

7. CONCLUSIONS

The study proposes a series of questionnaires to measure the users' perception of robots. This series will be called "Godspeed" because it is intended to help creators of robots on their development journey. Appendix A shows the application of the five Godspeed questionnaires using 5-point scales. It is important to notice that there is a certain overlap between anthropomorphism and animacy. The item artificial / lifelike appears in both sections. This is to be expected, since being alive is an essential part of being human-like.

When one of these questionnaires is used by itself in a study it would be useful to mask the questionnaire's intention by adding dummy items, such as optimistic / pessimistic. If multiple questionnaires are used then the items should be mixed so as to mask the intention. Before calculating the mean scores for anthropomorphism, animacy, likeability, or perceived intelligence it is good practice to perform a reliability test and report the resulting Cronbach's Alpha.

The interpretation of the results has, of course, some limitations. First, it is extremely difficult to determine the ground truth. In other words, it is complicated to determine objectively, for example, how anthropomorphic a certain robot is. Many factors, such as the cultural backgrounds of the participants, prior experiences with robots, and personality may influence the measurements. Taking all the possible biases into account would require a complex and therefore impracticable experiment. The resulting values of the measurements should therefore be interpreted not as absolute values, but rather as a tool for comparison. Robot developers can, for example, use the questionnaires to compare different configurations of a robot. The results may then help the developers to choose one option over the other. In the future, this set of questionnaires could be extended to also include the believability of a robot, the enjoyment of interacting with it, and the robot's social presence.

It is the hope of the authors that robot developers may find this collection of measurement tools useful. Using these tools would make the results in HRI research more comparable and could

therefore increase our progress. Interested readers, in particular experimental psychologists, are invited to continue to develop these questionnaires, and to validate them further.

A necessary development would be translation into different languages. Only native speakers can understand the true meanings of the adjectives in their language. It is therefore necessary to translate the questionnaires into the mother language of the participants. Appendix A includes the Japanese translation of the adjectives that we created using the back translation method. It is advisable to use the same method to translate the questionnaire into other languages. It would be appreciated if other translations are reported back to the authors of this study. They will then be collected and posted on this website:

http://www.bartneck.de/work/researchProjects/socialRobotics/godspeed

8. ACKNOWLEDGMENTS

The Intelligent Robotics and Communication Laboratories at the Advanced Telecommunications Institute International (Kyoto, Japan) supported this study.

9. REFERENCES

Bartneck, C., & Forlizzi, J. (2004). *A Design-Centred Framework for Social Human-Robot Interaction.* Proceedings of the Ro-Man2004, Kurashiki pp. 591-594. | DOI: 10.1109/ROMAN.2004.1374827

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2007). *Is the Uncanny Valley an Uncanny Cliff?* Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2007, Jeju, Korea pp. 368-373. | DOI: 10.1109/ROMAN.2007.4415111

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2008). My Robotic Doppelgänger – A Critical Look at the Uncanny Valley Theory. *Autonomous Robots*.

Bartneck, C., Kanda, T., Mubin, O., & Mahmud, A. A. (2007). *The Perception of Animacy and Intelligence Based on a Robot's Embodiment.* Proceedings of the Humanoids 2007, Pittsburgh.

Bartneck, C., & Rauterberg, M. (2007). HCI Reality - An Unreal Tournament. *International Journal of Human Computer Studies, 65*(8), 737-743. | DOI: 10.1016/j.ijhcs.2007.03.003

Bartneck, C., Verbunt, M., Mubin, O., & Mahmud, A. A. (2007). *To kill a mockingbird robot.* Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction, Washington DC pp. 81-87. | DOI: 10.1145/1228716.1228728

Berg, J. H., & Piner, K. (1990). Social relationships and the lack of social relationship. In W. Duck & R. C. Silver (Eds.), *Personal relationships and social support* (pp. 104-221). London: Sage.

Breemen, A., Yan, X., & Meerbeek, B. (2005). *iCat: an animated user-interface robot with personality.* Proceedings of the Fourth International Conference on Autonomous Agents & Multi Agent Systems, Utrecht. | DOI: 10.1145/1082473.1082823

Burgoon, J. K., & Hale, J. L. (1987). Validation and measurement of the fundamental themes for relational communication. *Communication Monographs, 54*, 19-41.

Chalmers, A. F. (1999). *What is this thing called science?* (3rd ed.). Indianapolis: Hackett.

Clark, N., & Rutter, D. (1985). Social categorization, visual cues and social judgments. *European Journal of Social Psychology, 15*, 105-119. | DOI: 10.1002/ejsp.2420150108

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489. | DOI: 10.1037/0022-0167.34.4.481

Dreyfus, H. L., & Dreyfus, S. E. (1992). *What computers still can't do : a critique of artificial reason*. Cambridge, Mass.: MIT Press.

Dreyfus, H. L., Dreyfus, S. E., & Athanasiou, T. (1986). *Mind over machine : the power of human intuition and expertise in the era of the computer*. New York: Free Press.

Fink, A. (2003). *The survey kit* (2nd ed.). Thousand Oaks, Calif.: Sage Publications.

Fogg, B. J. (2003). *Persuasive technology : using computers to change what we think and do*. Amsterdam ; Boston: Morgan Kaufmann Publishers.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*, 143-166. | DOI: 10.1016/S0921-8890(02)00372-X

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences, 40*(5), 873-884. | DOI: 10.1016/j.paid.2005.08.015

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243-249.

Inoue, K., Nonaka, S., Ujiie, Y., Takubo, T., & Arai, T. (2005). *Comparison of human psychology for real and virtual mobile manipulators.* Proceedings, IEEE International Conference on Robot and Human Interactive Communication pp. 73 - 78. | DOI: 10.1109/ROMAN.2005.1513759

Ishiguro, H. (2005). *Android Science - Towards a new cross-interdisciplinary framework.* Proceedings of the CogSci Workshop Towards social Mechanisms of android science, Stresa pp. 1-6.

Kiesler, S., & Goetz, J. (2002). *Mental models of robotic assistants.* Proceedings of the CHI '02 extended abstracts on Human factors in computing systems, Minneapolis, Minnesota, USA. | DOI: 10.1145/506443.506491

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of personality and social psychology 70*(1), 47-65. | DOI: 10.1037/0022-3514.70.1.47

Koay, K. L., Walters, M. L., & Dautenhahn, K. (2005). *Methodological Issues Using a Comfort Level Device in Human-Robot Interactions.* Proceedings of the IEEE RO-MAN pp. 359 - 364.

Koda, T. (1996). *Agents with Faces: A Study on the Effect of Personification of Software Agents.* Master Thesis, MIT Media Lab, Cambridge.

Kulic, D., & Croft, E. (2005). *Anxiety Detection during Human-Robot Interaction.* Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Edmonton, Canada pp. 389 - 394. | DOI: 10.1109/IROS.2005.1545012

Kulic, D., & Croft, E. (2006). *Estimating Robot Induced Affective State Using Hidden Markov Models.* Proceedings of the RO-MAN 2006 – The 15th IEEE International Symposium on Robot and Human Interactive Communication, Hatfield pp. 257-262. | DOI: 10.1109/ROMAN.2006.314427

Lee, K. M., Park, N., & Song, H. (2005). Can a Robot Be Perceived as a Developing Creature? *Human Communication Research, 31*(4), 538-563. | DOI: 10.1111/j.1468-2958.2005.tb00882.x

Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J. (2001). A cross-media presence questionnaire: The itc sense of presence inventory. *Presence: Teleoperators and Vir tual Environments, 10*(3), 282-297. | DOI: 10.1162/105474601300343612

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*.

MacDorman, K. F. (2006). *Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley.* Proceedings of the ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science, Vancouver.

McAleer, P., Mazzarino, B., Volpe, G., Camurri, A., Patterson, H., & Pollick, F. (2004). Perceiving Animacy and Arousal in Transformed Displays of Human Interaction. *Journal of Vision, 4*(8), 230-230. | DOI: 10.1167/4.8.230

Minato, T., Shimada, M., Itakura, S., Lee, K., & Ishiguro, H. (2005). *Does Gaze Reveal the Human Likeness of an Android?* Proceedings of the 4th IEEE International Conference on Development and Learning, Osaka. | DOI: 10.1109/DEVLRN.2005.1490953

Monathan, J. L. (1998). I Don't Know It But I Like You - The Influence of Non-conscious Affect on Person Perception. *Human Communication Research, 24*(4), 480-500. | DOI: 10.1111/j.1468-2958.1998.tb00428.x

Nass, C., & Reeves, B. (1996). *The Media equation*. Cambridge: SLI Publications, Cambridge University Press.

Nonaka, S., Inoue, K., Arai, T., & Mae, Y. (2004). *Evaluation of Human Sense of Security for Coexisting Robots using Virtual Reality.* Proceedings of the IEEE International Conference on Robotics and Automation, New Orleans, LA, USA pp. 2770-2775. | DOI: 10.1109/ROBOT.2004.1307480

Nunnally, J. C. (1978). *Psychometric theory* (2d ed.). New York: McGraw-Hill.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurements of meaning*. Champaign: University of Illinois Press.

Parise, S., Kiesler, S., Sproull , L. D., & Waters , K. (1996). *My partner is a real dog: cooperation with social agents.* Proceedings of the 1996 ACM conference on Computer supported cooperative work, Boston, Massachusetts, United States pp. 399-408. | DOI: 10.1145/240080.240351

Powers, A., & Kiesler, S. (2006). *The advisor robot: tracing people's mental model from a robot's physical attributes.* Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, Salt Lake City, Utah, USA. | DOI: 10.1145/1121241.1121280

Rani, P., Sarkar, N., Smith, C. A., & Kirby, L. D. (2004). Anxiety detecting robotic system - towards implicit human-robot collaboration. *Robotica, 22*, 85-95. | DOI: 10.1017/S0263574703005319

Rani, P., Sims, J., Brackin, R., & Sarkar, N. (2002). Online stress detection using phychophysiological signals for implicit human-robot cooperation. *Robotica, 20*(6), 673-685. | DOI: 10.1017/S0263574702004484

Robbins, T., & DeNisi, A. (1994). A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology, 79*, 341-353. | DOI: 10.1037/0021-9010.79.3.341

Sarkar, N. (2002). *Psychophysiological Control Architecture for Human-Robot Coordination - Concepts and Initial Experiments.* Proceedings of the IEEE International Conference on Robotics and Automation, Washington, DC, USA pp. 3719-3724. | DOI: 10.1109/ROBOT.2002.1014287

Scholl, B., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*(8), 299-309. | DOI: 10.1016/S1364-6613(00)01506-0

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*(3), 417-457.

Sony. (1999). Aibo. Retrieved January, 1999, from http://www.aibo.com

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception, 29*(8), 943-951. | DOI: 10.1068/p3101

Wada, K., Shibata, T., Saito, T., & Tanie, K. (2004). Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proceedings of the IEEE, 92*(11), 1780-1788. | DOI: 10.1109/JPROC.2004.835378

Warner, R. M., & Sugarman, D. B. (1996). Attributes of Personality Based on Physical Appearance, Speech, and Handwriting. *Journal of Personality and Social Psychology, 50*(4), 792-799. | DOI: 10.1037/0022-3514.50.4.792

Weizenbaum, J. (1976). *Computer power and human reason : from judgment to calculation*. San Francisco: W. H. Freeman.

Appendix A: Overview of the Godspeed Questionnaire series using a 5-point scale.

GODSPEED I: ANTHROPOMORPHISM

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Fake 偽物のような	1	2	3	4	5	Natural 自然な
Machinelike 機械的	1	2	3	4	5	Humanlike 人間的
Unconscious 意識を持たない	1	2	3	4	5	Conscious 意識を持っている
Artificial 人工的	1	2	3	4	5	Lifelike 生物的
Moving rigidly ぎこちない動き	1	2	3	4	5	Moving elegantly洗練された動き

GODSPEED II: ANIMACY

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Dead 死んでいる	1	2	3	4	5	Alive 生きている
Stagnant 活気のない	1	2	3	4	5	Lively 生き生きとした
Mechanical 機械的な	1	2	3	4	5	Organic 有機的な
Artificial 人工的な	1	2	3	4	5	Lifelike 生物的な
Inert 不活発な	1	2	3	4	5	Interactive 対話的な
Apathetic 無関心な	1	2	3	4	5	Responsive 反応のある

GODSPEED III: LIKEABILITY

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Dislike 嫌い	1	2	3	4	5	Like 好き
Unfriendly 親しみにくい	1	2	3	4	5	Friendly 親しみやすい
Unkind 不親切な	1	2	3	4	5	Kind 親切な
Unpleasant 不愉快な	1	2	3	4	5	Pleasant 愉快な
Awful ひどい	1	2	3	4	5	Nice 良い

GODSPEED IV: PERCEIVED INTELLIGENCE

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Incompetent 無能な	1	2	3	4	5	Competent 有能な
Ignorant 無知な	1	2	3	4	5	Knowledgeable 物知りな
Irresponsible 無責任な	1	2	3	4	5	Responsible 責任のある
Unintelligent 知的でない,	1	2	3	4	5	Intelligent 知的な
Foolish 愚かな	1	2	3	4	5	Sensible 賢明な

GODSPEED V: PERCEIVED SAFETY

Please rate your emotional state on these scales:
以下のスケールに基づいてあなたの心の状態を評価してください。

Anxious 不安な	1	2	3	4	5	Relaxed 落ち着いた
Agitated 動揺している	1	2	3	4	5	Calm 冷静な
Quiescent 平穏な	1	2	3	4	5	Surprised 驚いた

Social resonance: a theoretical framework and benchmarks to evaluate the social competence of humanoid robots

Thierry Chaminade
CNRS UMR 6193
31, Chemin Joseph Aiguier
13402 Marseille, FRANCE
(33) 4 91 16 45 38
tchamina@gmail.com

ABSTRACT
As artificial anthropomorphic agents such as humanoid and android robots are increasingly present in our societies, it is important to understand humans' automatic and unconscious reactions to these agents. "Social resonance" is an emerging framework in social cognitive neuroscience, based on the finding of an overlap between cognitive processes used when experiencing a mental state and when perceiving another individual experiencing the same mental state. It has been applied to the domains of action, emotion and pain. After presenting this framework and discussing its use to address questions pertaining to artificial agents' social competence I will present two types of benchmark tests that have been used to test social resonance effects of humanoid robots. A first type uses paradigms derived from experimental psychology to investigate humans' responses to artificial agents. Social resonance, applied to the domain of action, implies interference between observed and executed actions, which can be measured as a function of whether the observed agent is a robot or a human. A second series of works in progress uses human functional neuroimaging to investigate the brain response to the observation of humanoid robots actions or emotions. Activity in regions resonating to the observation of humans is probed against robots to compare their responses to both agents.

Categories and Subject Descriptors
H5.2 [**Information Interfaces and Presentation**]: User Interfaces - *User-centered design*; *Interaction styles*; *Theory and methods.*

General Terms
Experimentation.

Keywords
Social cognitive neuroscience, resonance, mirror neurons, humanoids.

1. INTRODUCTION
Artificial anthropomorphic agents such as humanoid and android robots are increasingly present in our societies. Aichi 2005 exposition broadcasted internationally tens of robots, from task-specialized robots to social androids welcoming the visitors. Everyday use of robots is becoming accessible, as with the example of Kokoro's company simroid, a feeling and responsive android patient for use as a training tool for dentists, or robotic companions being introduced for use with children [1] or elderly people.

For these robots to interact optimally with humans, it is important to understand humans' automatic and unconscious reactions to these agents. Studies have addressed the issue of the form [2] and functionalities [3-6] a humanoid robot should have in order to be socially accepted. Both types of approaches have mostly relied on introspective judgments or implicit assumptions, such as the need for human traits, which may bias their conclusions. Conversely, the "Uncanny Valley of eeriness" hypothesis proposes that artificial agents imperfectly attempting to impersonate humans induce a negative emotional response [7, 8]. The felt creepiness of robots such as the simroid tends to confirm this hypothesis, which has served for years as a guideline to avoid realistic anthropomorphism in robotic designs. Take for example Toshitada Doi, then Sony's corporate executive vice president, on the design of qrio, Sony's humanoid robot: *"We suggested the idea of an "eight year-old space life form" to the designer -- we didn't want to make it too similar to a human. In the background, as well, lay an idea passed down from the man whose work forms the foundation of the Japanese robot industry, Masahiro Mori: "the valley of eeriness". If your design is too close to human form, at a certain point it becomes just too . . . uncanny. So, while we created QRIO in a human image, we also wanted to give it little bit of a "spaceman" feel."* Nowadays though, people like David Hanson, founder of Hanson robotics, builds realistic anthropomorphic robots under the assumption that the uncanny valley is an illusion caused by the poor quality of aesthetic designs [9]. Its introduction of an elastic polymer efficiently mimicking the human skin, Frubber, has improved the subjective quality of his robots, an argument in favor of the idea that if it exists, the uncanny valley is everything but insurmountable.

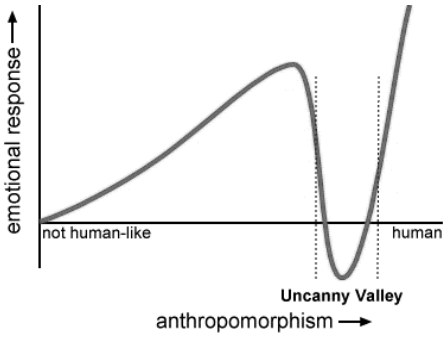


Figure 1: Sketch of the uncanny valley. The Valley represents the negative emotional response hypothesized by Masahiro Mori in response to realistic albeit imperfect anthropomorphic agent.

Despite its importance in robotic design, and, more widely, in the robotic industry, the theory of the uncanny valley has only recently started to be investigated with scientific tools. Karl MacDorman and Iroshi Ishiguro in particular, through their collaboration with the robotic firm Kokoro in the development of increasingly acceptable androids, have played a major role in naturalizing the investigation of the uncanny valley [8]. Yet Masahiro Mori's hypothesis is increasingly being seen as unpractical. While it is possible to describe human emotional reactions along one axis, restricting it to its valence for instance – positive versus negative emotions -, such a reduction of dimensionalities is deeply unsatisfying when applied to reduce robotic designs anthropomorphism along one axis. Despite a number of laudable attempts, more particularly using morphs between real and artificial agents, the linearity of "anthropomorphism" as a single measure remains dubious.

As a consequence, these questions about robotic designs, not limited to the uncanny valley but extended to all features that modulate the social competence of artificial agents, should be firmly grounded in a theoretical framework and follow robust experimental paradigms. I will present how such a strategy is being implemented using the "social resonance" hypothesis, which offers plausible explanations of implicit aspects of social interactions between human agents in everyday life. An agent is an entity able to produce an action, in other words, to have a perceivable effect on the world. Classically, agents refer to intentional (or real) agents, that is living creatures and prominently humans, for which intentionality of behavior can be assessed directly. Treating robots as "artificial agents", the social resonance hypothesis can be applied to the investigation of human reactions towards robotics designs. "Artificial agent" here refers to any artificial, man-made entity that resembles an intentional agent in that it produces seemingly intentional actions. This definition includes humanoid robots comprising the realistic androids, but also computer animations of anthropomorphic characters. Social competence is their ability to engage in natural exchanges, or social interactions, with intentional agents.

In a first part, I will introduce the theoretical framework of social resonance. In the second part, I will give specific examples of how hypotheses derived from this framework can be tested experimentally. In a third part, I will introduce more recent attempts to probe social resonance in response to artificial agents using human functional neuroimaging.

2. SOCIAL RESONANCE: A THEORETICAL FRAMEWORK

2.1 Overlook

Following the finding that the same neural structures show an increase of activity both when executing a given action and when observing another individual executing the same action, theories of social behaviours using concepts of resonance have flourished in the scientific literature [10-12]. Similar ideas can be traced back as far as William James in the 19th century [13], or more recently the theory of event coding [14]. In the domain of action, neuropsychological findings hinted, in the early 1990s, that gesture perception and limb praxis share the same cortical circuits [15]. Similarly in language, the motor theory of speech perception claimed, on the basis of experimental data, that the object of speech perception are not sounds, but the phonetic gestures of the speaker, whose neural underpinnings are motor commands [16]. I refer to this process as motor resonance, which is defined, at the behavioural and neural levels, as the automatic activation of motor control systems during perception of actions. Mirror neurons renewed interest in these processes by offering the first demonstration that resonance had validity at the cellular level.

2.2 Neurophysiology of resonance

2.2.1 Macaque monkey mirror neurons

Mirror neurons are a type of neuron found in the macaque monkey brain and defined by their response, as recorded by single cell electrophysiological recordings. First reported in 1992 by Giacomo Rizzolatti's group in Parma [17], they were officially named "mirror neurons" in a 1996 Cognitive Brain Research report as "a particular subset of F5 neurons [which] discharge[s] when the monkey observes meaningful hand movements made by the experimenter" [18]. The importance of this discovery stems from the known function of area F5, a premotor area in which neurons discharge when monkeys execute distal goal-directed motor acts such as grasping, holding or tearing an object. Comparing the various reports, it is reasonable to assume that around 20% of recordable neurons in these areas have mirror properties in a loose sense, but only a lower percentage, around 5%, shows action specificity (i.e. the same action is the most efficient in causing the neuron to fire when the monkey observes and when he executes it).

2.2.2 Human neuroimaging data: action observation

The human physiological data, using the brain imaging techniques which emerged in the last decades such as positron emission tomography (PET), functional magnetic resonance imagery (fMRI), electroencephalography (EEG), magnetoencephalography (MEG) and transcranial magnetic stimulation (TMS), entails an expected conclusion on the basis of the mirror neuron literature in macaque monkey: premotor cortices, originally considered to be exclusively concerned with motor control, are also active during observation of actions in the absence of any action execution [19]. What remains unknown is whether the same brain region, and a fortiori the same neurons, would be activated by the observation and the execution of the same action in the whole of the premotor system, or whether this specificity is limited to a small percentage of ventral premotor neurons. In other words, are all premotor regions activated in response to the observation of

action populated with mirror neurons? But irrespective of the answer to this question, accumulating human neuroimaging data does confirm in humans what mirror neurons demonstrated beyond doubt in macaque monkeys at the cellular level: neurophysiological bases for the perception of other individuals' behaviors makes use of the neurophysiological bases for the control of the self's behavior.

2.2.3 Human neuroimaging data: generalization of resonance

An intriguing recent trend in the human literature is that this resonance is not limited to observation of object-directed hand actions, as mirror neurons are, but generalizes to a number of other domains of cognition. For example, an fMRI study investigated touch perception by looking for overlap between being touched and observing someone being touched [20]. An overlap of activity was found in the secondary somatosensory cortex, a brain region involved in integrating somatosensory information with other sensory modalities such as touch. Another study reported activity in the primary sensory cortex during the observation of touch [21]. Thus, there is a resonance for touch, by which observation of someone else being touched recruits neural underpinnings of the feeling of touch. In the same vein, observation of the expression of disgust activates a region of the insula also activated during the feeling of disgust caused by a nauseating smell [22]. A neuron in the anterior cingulate cortex, which participates in pain perception, fired when a patient experienced pinpricks and when he observed the examiner receiving the same painful stimulus [23]. A last example concerns emotions: the amygdala, a brain structure involved in feeling of primary emotions such as fear is fundamental in recognizing fear from facial expressions [24].

The mirror neurons studied in macaque monkey are a very specific example of a more general mechanism of human cognition, namely the fact that neuronal structures used when we experience a mental state, including but not limited to internal representation of an action, are also used when we perceive other individuals experiencing the same mental state. These examples support a generalization of motor resonance to other domains of cognition such as emotions and pain, that can be transferred between interacting agents, hence the term of social resonance.

2.3 Functions of resonance in social interactions

Motor resonance is evident in behaviors like action contagion (contagion of yawning for example), motor priming (the facilitation of the execution of an action by seeing it done [25]) and motor interference (the hindering effect of observing incompatible actions during execution of actions [26]). But, does the motor resonance described in a laboratory environment have a significant impact in everyday life? The chameleon effect was introduced to describe the unconscious reproduction of "postures, mannerisms, facial expressions and other behaviors of one's interacting partner" [27]. This effect can easily be experienced in face-to-face interactions, when one crosses his arms or legs to see his partner swiftly adopt the same posture. Subjects unaware of the purpose of the experiment interacted with an experimenter performing one of two target postures, rubbing the face or shaking the foot. Analysis of the behavior showed a significant increase of the tendency to engage in the same action. In addition this imitation makes the person facing considered as more likable even though you are not aware of this imitation [27]. This mimicry has been described as a source of empathy [28], and motor resonance offers a parsimonious system to automatically identify with conspecifics.

The main function classically attributed to resonance is action understanding. The most convincing argument to date comes from neuropsychology, the study of cognitive impairments consecutive to brain lesions. It was recently reported that premotor lesions impair the perception of biological motion presented using point-light displays [29]. Therefore, not only are premotor cortices activated during the perception of action, but also their lesion impairs the perception of biological motion, demonstrating that they have are functionally involved in the perception of action.

Another function frequently associated with resonance is imitation. Imitation covers a continuum of behaviors ranging from simple, automatic and involuntary action contagion to intentional imitation and emulation [30]. Jacobs and Jeannerod recently emphasized that "[imitation] is a folk psychology concept whose boundaries are presently too ill-defined for scientific purposes" [31]. It is difficult to realize the number of complex mechanisms involved in imitation, from body correspondence to extraction of task-relevant features [32]. Yet, key regions for the human imitation are the left inferior parietal lobule [33], possible homolog of the macaque monkey area PF, and the ventral premotor cortex [34], homolog of the macaque monkey area F5, both being the regions in the macaque monkey brain where mirror neurons were reported.

More recently, resonating systems have been found in other domains such as empathy for pain [35], disgust [22], and led to the hypothesis that this generalized resonance between oneself and other selves, or social resonance, underlies a number of social behaviors such as imitation [12], action understanding [36], social bonding and empathy. Social resonance is central to the understanding of social behaviors [37], and the methods used to investigate it should be extended to the measure of the social competence of anthropomorphic artificial agents including robots. The underlying assumption is that the measure of resonance will indicate the extent to which an artificial agent is considered as a social inter-actor.

3. A BEHAVIORAL MEASURE OF SOCIAL RESONANCE: MOTOR INTERFERENCE

3.1 Canonical paradigm

A consequence of motor resonance, motor interference is the influence the perception of another individual's actions has on the execution of actions by the self: observing an action facilitates the execution of the same action, and hinders the execution of a different action.

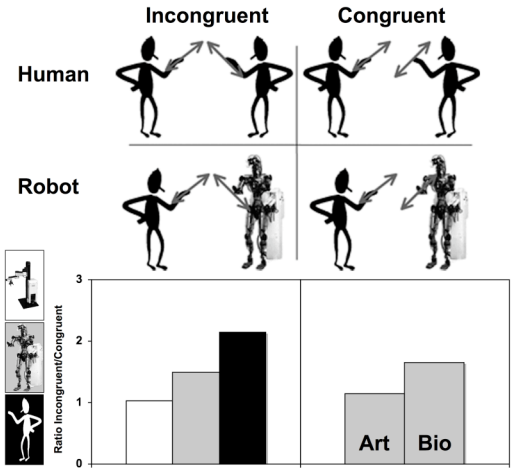


Figure 2: top: factorial plan showing the 4 canonical condition of motor interference experiment: horizontally, the spatial congruency between the volunteers and the tested agent movement; vertically, the human control and the agent being tested, in this case the humanoid robot DB. Bottom: summary of the experimental result. The ratio between the variance for incongruent and congruent movements is shown for three agents, an industrial robot [26], a humanoid robot [38] and a human [26, 38] on the left, and for the humanoid robot acting with artificial (ART) or biological (BIO) motion [39] on the right.

In one of the experimental paradigms developed to investigate motor interference, volunteers were asked to raise their fingers in response either to a symbolic cue appearing on a nail or to a movement of the finger of a hand presented visually [40]. The two cues could be present on the same finger (congruent cues) or on different fingers (incongruent cues). In the later case, there were two conflicting cues and only one was relevant for the volunteers. It was found that the observation of an incongruent finger movement hindered the response to the symbolic cue –increase of the time needed to respond- but that the reverse effect (symbolic cue hindering the response to the finger movement) was very small. In other word, when responding to a symbolic cue, the response is hindered by the observation of an incompatible action and facilitated by a compatible one. In this paradigm, producing an action similar to an observed action is a prepotent response that requires to be inhibited to execute the correct response.

3.2 Application to the study of robotic designs

A series of experiments was initiated by Kilner et al's [26] study of motor interference when facing a real human being or an industrial robotic arm. Volunteers in this study produced a vertical or horizontal arm movement while watching another agent in front of them producing a spatially congruent (ie vertical when vertical, horizontal when horizontal) or a spatially incongruent (horizontal when vertical and vertical when horizontal) movement. The interference effect was measured by the increase of the variance of a movement, was found when subjects watched an arm movement spatially incompatible with the one they were producing (e.g. vertical versus horizontal, Figure 2) [26]. Interestingly, Kilner et al.'s study did not find any interference effect using an industrial robotic arm moving at a constant velocity, suggesting at first that motor interference was specific to interactions between human agents.

3.2.1 Effect of form

This experimental paradigm was adapted to investigate how humanoid robots interfere with humans. Subjects performed rhythmic arm movements while observing either a human agent or humanoid robot performing either congruent or incongruent movements with comparable kinematics. The variance of the executed movements was used as a measure of motor interference caused by the observed action. We found that in contrast to the industrial robotic arm, a humanoid robot executing movements based on motion captured data caused a significant change of the variance of the movement depending on congruency [38]. The ratio between the variance in the incongruent and in the congruent conditions increases from the industrial robotic arm (r=1, no increase in incongruent condition), the humanoid robot (r~1.5) and the human (r~2), both in our and in Kilner et al., experiment [26].

3.2.2 Effect of motion

In a follow-up experiment, we investigate the effect of the movement kinematics on the interference. The humanoid robot moved either with a biological motion based, as previously, on recorded trajectories, or with an artificial motion implemented by a 1-DOF sinusoidal movement of the elbow. We found a significant effect of the factors defining the experimental conditions. The increase in incongruent conditions was only significant when the robot movements followed biological motion [39]. The ratio that could be calculated on the basis of the results was, in the case of biological motion, comparable to the ratio reported in the previous experiment, ~1.7. Note the importance of having internal controls, in this case human agents, to compare the ratio within groups.

3.2.3 Effect of visibility

Another factor capable of influencing motor resonance has been tested recently, and its analysis is still in progress. The effect of interference could be due merely to the appearance of the agent, which would predict a linear increase of the ratio between the variance for incongruent and congruent movements with anthropomorphism. Alternatively it could be influenced by the knowledge we have about the nature of the other agent. At the level of brain physiology, it is known that thinking we interact with another individual or with a computer algorithm changes local brain activity despite the fact that the same algorithm is actually controlling the interaction [41].

To test whether appearance was the main factor we covered the body and face of both agents, the human and the humanoid robot, with a black cloth leaving just the arm visible, and compared the results of the interference paradigm between covered and uncovered agents. Preliminary results indicate that the variance is increased in all conditions, implying that motor interference can be measured in the absence of full body visibility and suggesting that knowledge about the aspect of the agent being interacted with is sufficient to elicit motor resonance (bottom-down effect of the knowledge). Alternatively arm movements, from either a human or a humanoid robot, could display sufficient cues about the nature of the agent being interacted with to elicit motor resonance (top-down effect of the stimulus). Further analysis of the data, investigating the effect of the nature and of the visibility of the agents on the interference effect is still required to unravel the two possible explanations. A different but comparable

experiment derived from the motor resonance hypothesis, motor priming, has provided results in favor of the second hypothesis, a bottom-up effect due to the appearance of the robotic device.

3.2.4 Motor priming with a robotic hand

A very similar result has been obtained using motor priming. Motor priming can be conceptually conceived as a consequence of motor resonance opposite to the motor interference measured in the previous part. Its foundation is that observing an action facilitates ("primes") the execution of the same action, and can be described as "automatic imitation". Responses are faster and more accurate when they involve executing the same movement than executing another movement.

This effect was investigating with two actions, hand opening and hand closing, in response to the observation of a hand opening and closing, with the hand being either a realistic human hand or a simple robotic hand having the appearance of an articulated claw with two opposite fingers [42] (see Figure 3, top right). Volunteers in the experiment were required to make a prespecified response (to open or to close their right hand) as soon as a stimulus appears on the screen. Response time was recorded and analyzed as a function of the content of the stimulus, either a human or a robotic hand, in a posture congruent or incongruent with the prespecified movement (eg open or closed hand when the prespecified action is opening the hand).

Results showed an increased response time in incongruent compared to congruent conditions, in response to both human and robotic hand, suggesting that the motor priming effect was not restricted to human stimuli but generalized to robotic stimuli [42]. As with the motor interference measure, the size of the effect, taking the form of the time difference between response to incongruent and congruent stimuli, was larger for human stimuli (~30 ms) that for robotic stimuli (~15 ms).

A follow-up experiment tested whether the effect is better explained by a bottom-up process due to the overall shape or a top-down process caused by the knowledge of the intentionality of humans compared to robotic devices [43]. Human hands were modified by the addition of a metal and wire wrist, and were perceived as less intentional than the original hands. Nevertheless in the priming experiment, no significant differences were found between the priming effect of the original and of the robotized human hand, in favor of the bottom-up hypothesis that the overall hand shape, and not its description as a human or robotic hand, affects the priming effect.

Overall, these accumulating results confirm the validity of using motor interference as a metric of motor resonance, a possible proxy for social competence, with humanoid robots. First, motor resonance is an important aspect of social cognition, particularly important in automatic and unconscious perception of other agents. Second, the effects of motor interference on behavior can be measured easily, as movement variance or reaction time. Third, existing results strongly suggest the effect is modulated by the appearance of the agent being tested. And finally, these interference effects have been shown to increase with the realism of the stimulus.

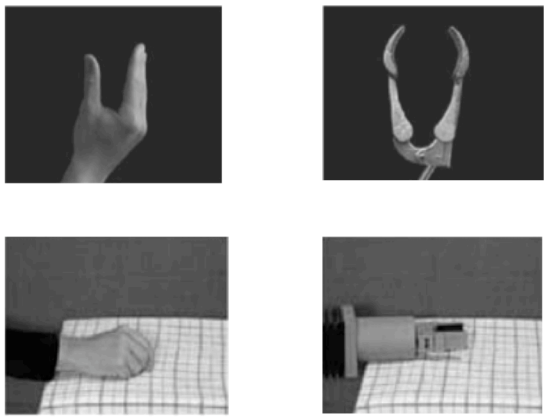


Figure 3: human (left) and robotic (right) hand stimuli used (top) in the motor priming experiment described in 3.2.4 [42] and (bottom) in the fMRI experiment investigating perception of hand actions described in 4.1.2 [44]. Images adapted from the original publications.

4. FUNCTIONAL NEUROIMAGING OF SOCIAL RESONANCE

Motor resonance is not only a behavioral observation, but also, and in some ways foremost, a neuroscience result, for which interest was largely renewed by the discovery of mirror neurons (see 2.2.1) in monkeys and the subsequent finding of an overlap between brain activity in response to experiencing states and to observing other individuals experiencing similar states (see 2.2.2 and 2.2.3). Thus measures of changes in brain activity in regions specific to motor, or emotional, resonance, depending on whether an observed action is executed by a real human, a realistic android or a humanoid robot, can provide an objective measure of the resonance between the observer and the observed agent. I will present recent attempts here.

4.1 Perception of hand actions

The first experiments focused on the observation of hand actions. Two independent experiments came to conflicting results about activity in the human premotor region thought to be homologous to the region in which mirror neurons were identified in macaque monkeys and known to be active during the observation of actions.

4.1.1 PET experiment

In the first experiment, volunteers observed a human hand or a robotic hand grasping objects while having their brain scanned by PET [45]. Results indicate that activity in a region of interest in the human premotor cortex, based on the assumption that this pat of the human cortex contains mirror system, failed to report any significant activity in response to the robotic moving compared to static stimuli. In contrast they reported strong activity in early visual areas of the lateral occipital cortex.

4.1.2 fMRI experiment

To test whether the resonance system responding to human action would respond to a robotic action, another group of researchers [44] localized motor resonance areas by looking for an overlap between the areas involved in motor execution and areas responding to the observation of movie clips of object-directed actions depicted by a human or by a robotic hand. The robotic hand, in this case, consisted of a simple claw with two

sides allowing a simple grasp (see Figure 3, bottom right).
Regions of the premotor cortex, both ventral and dorsal, and of the parietal cortex were activated strongly by the sight of both human and robotic complex actions compared to simple movement or static images, with no significant differences between these two agents [44].

These two results appear contradictory, but a number of differences between the two experiments could explain this discrepancy. The difference of neuroimaging methodology could be responsible (PET and fMRI), as well as differences in experimental paradigm; the use of repeated stimuli in the PET study was advanced in [44]. Another possibility, that has not been proposed but would fit the results presented in the previous section and the thesis defended here, would be that the resonance system would respond differently because of the difference between the appearance of the robotic hand, and in particular their anthropomorphism. Such an interpretation provides a fascinating hypothesis to further investigate the response of the neural bases of motor resonance to robotic designs.

4.2 Perception of emotions

4.2.1 Presentation of the experiment

In a collaborative work between the university of Pisa in Italy (Maria-Alessandra Umilta, Vittorio Gallese, Gicomo Rizzolatti), Waseda University in Japan (Atsuo Takanishi, Massimiliano Zecca) and University College London in United Kingdom (Thierry Chaminade, Sarah-Jayne Blakemore, Chris Frith), we recorded the fMRI response to emotions depicted by a a human or the humanoid robot. The humanoid robot Waseda Eye No.4 Refined II (WE-4RII, [46]) is the result of the integration of robotic hands into an upper body emotional robot. WE-4RII being designed to express human emotions, it contains 26 degrees of freedom for facial and head movements (including 4 for the neck and 8 for the eyebrows). It can express its emotion using facial expression, neck and waist as well as arms and hands motion. Both the posture and the motion velocity are controlled to realize the effective emotional expressions.

Stimuli consisted of 1.5-seconds greyscale videoclips depicting either one of 3 emotions (Joy, Anger and Disgust) or the emotionally neutral Silent Speech. All stimuli started from a neutral pose and stopped with the emotional expression (Figure 4). Great care was taken to match the dynamics of the human and robot stimuli pairwise. Two tasks were presented to the experiment volunteers, judging the emotionality of the stimulus or the quantity of motion.

Main question was whether the regions involved in motor and in emotional resonance, the later depending on the emotion (e.g. amygdala for anger and anterior insula for disgust) would respond to both types of agents, and with the same intensity. In other words, can we find at the brain level a reduced measure of resonance for the robot compared to the real human.

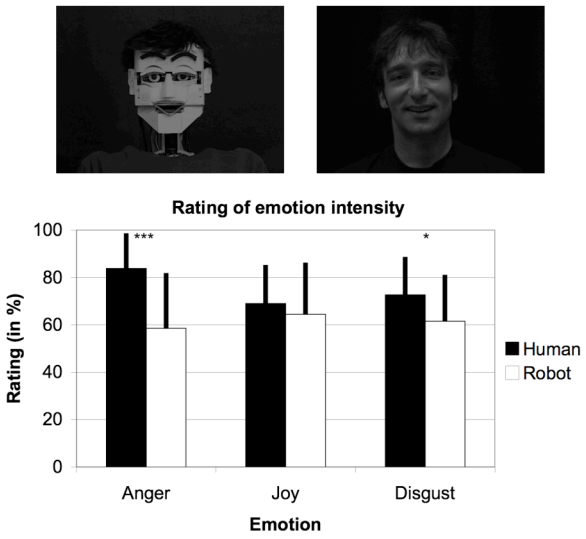


Figure 4: top: examples of happy faces from the humanoid robot (left) and the human (right); bottom: percentage of emotional ratings (mean + standard error; rating scale) for the three emotions and the two agents. Ratings are significantly higher for the human in the case of anger (*: $p<0.001$) and of disgust (*: $p<0.05$).**

4.2.2 Results

Only preliminary observations from this unpublished work can be discussed at this time. First, while all stimuli were considered as emotional, human stimuli were judged as significantly more emotional for Anger and Disgust, but not for Joy (Figure 4). As a consequence, the difference in perceived emotionality is unlikely to explain fully the brain imaging results.

The brain activity can be summarized as follows. Taken together, there is no activity in the motor resonance system when emotional stimuli are compared to neutral stimuli across all agents and tasks, which could be explained by the fact that both the target emotional stimuli and the control stimuli contain upper torso actions which are removed by the subtraction. When the effect of human and robotic agents were contrasted, brain responses to the humanoid robot were located in the posterior part of the brain, dedicated to the perception of objects, similar to the region already discussed in 4.1.1. In contrast, increased brain responses to the human were found in cognitively higher regions, and in particular, in regions specific to the perception of the three emotions, in the amygdala for Joy and Anger and in the insula for Disgust.

While more analysis is needed to refine these results they do point to an unexpected result. Even when the stimuli are judged consciously as equally emotional, the brain response indicates that emotional resonance is restricted to the observation of human stimuli. Thus, as long as emotions are concerned, our results do not confirm an increased resonance with anthropomorphic robots.

5. CONCLUSIONS

Because it is a solid theoretical basis from social cognitive neuroscience, social resonance is a promising framework to

describe interactions between natural and artificial agents. Behavioral results, based on experimental paradigm originally developed to investigate human social behavior, indicate that motor resonance is modulated by the appearance of the agent, with intermediate measure for anthropomorphic robotic devices in comparison to non anthropomorphic devices and real humans. It offers a rich ensemble of objective benchmarks to measure the social competence of artificial agents. Neuroimaging also promises to be a rich tool, though at present there is no clearly defined experimental paradigm shared by the community. Further work is required to develop more efficient paradigm, but such approaches present the advantage of relying on a large amount of existing results on the resonance system in humans. The experiments described here can be easily reproduced by other roboticists to test the social competence of their robots, having the potential of providing a unifying measure allowing comparisons of various robotic designs for humanoids.

6. REFERENCES

[1] Tanaka, F., Cicourel, A., and Movellan, J. R., 2007. Socialization between toddlers and robots at an early childhood education center, Proceedings of the National Academy of Sciences, vol. 104, pp. 17954-17958.

[2] Disalvo, C., Gemperle, F., Forlizzi, J., and Kiesler, S., 2002. All robots are not created equal: the design and perception of humanoid robot heads, Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques, vol. http://doi.acm.org/10.1145/778712.778756.

[3] Scassellati, B., 2001. Foundations for a theory of mind for a humanoid robot, B. Scassellati, Foundations for a theory of mind for a humanoid robot, Ph.D. Thesis, Dept. Elec. Eng. Comp. Sci., MIT, 2001.

[4] Kozima, H. and Yano, H., "A robot that learns to communicate with human caregivers," presented at 1st International Workshop on Epigenetic and Robotics, Lund, Sweden, 2001.

[5] Kozima, H. and Yano, H., 2001. A robot that learns to communicate with human caregivers, H. Kozima and H. Yano, A robot that learns to communicate with human caregivers, in: Proc. Intl. Wksp. Epigenetic Rob., 2001.

[6] Breazeal, C. and Scassellati, B., 2000. Infant-like Social Interactions between a Robot and a Human Caretaker, CaretakerAdaptive Behavior, vol. 8, pp. 49-74.

[7] Mori, M., 1970. The valley of eeriness (japanese), Energy, vol. 7, pp. 33-35.

[8] Macdorman, K. F. and Ishiguro, H., 2006. The uncanny advantage of using androids in cognitive and social science research, Interaction Studies, vol. 7, pp 297-338.

[9] Hanson, D., "Expanding the Aesthetics Possibilities for Humanlike Robots," presented at Proc. IEEE Humanoid Robotics Conference, special session on the Uncanny Valley, Tskuba, Japan, 2005.

[10] Blakemore, S. J. and Decety, J., 2001. From the perception of action to the understanding of intention, Nature Reviews Neuroscience, vol. 2, pp. 561-567.

[11] Gallese, V., Keysers, C., and Rizzolatti, G., 2004. A unifying view of the basis of social cognition, Trends in Cognitive Sciences, vol. 8, pp. 396-403.

[12] Rizzolatti, G., Fogassi, L., and Gallese, V., 2001. Neurophysiological mechanisms underlying the understanding and imitation of action, Nature Reviews Neuroscience, vol. 2, pp. 661-670.

[13] James, W., *Principles of Psychology*. New York: Holt, 1890.

[14] Hommel, B., Musseler, J., Aschersleben, G., and Prinz, W., 2001. The Theory of Event Coding: A Framework for Perception and Action Planning, Behav Brain Sci, vol. 24, pp. 849-937.

[15] Rothi, L. J. G., Ochipa, C., and Heilman, K. M., 1991. A cognitive neuropsychological model of limb praxis, Cognitive Neuropsychology, vol. 8, pp. 443-458.

[16] Liberman, A. M. and Mattingly, I. G., 1985. The motor theory of speech perception revised, Cognition, vol. 21, pp. 1-36.

[17] Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G., 1992. Understanding motor events: a neurophysiological study, Exp Brain Res, vol. 9, pp. 176-180.

[18] Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L., 1996. Premotor cortex and the recognition of motor actions, Cognitive Brain Research, vol. 3, pp. 131-141.

[19] Chaminade, T. and Decety, J., 2001. A common framework for perception and action: neuroimaging evidence, Behav Brain Sci, vol. 24, pp. 879-882.

[20] Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L., and Gallese, V., 2004. A touching sight: SII/PV activation during the observation and experience of touch, Neuron, vol. 42, pp. 335-46.

[21] Blakemore, S. J., Bristow, D., Bird, G., Frith, C., and Ward, J., 2005. Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia, Brain, vol. 128, pp. 1571-83.

[22] Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., and Rizzolatti, G., 2003. Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust, Neuron, vol. 40, pp. 655-64.

[23] Hutchison, W. D., Davis, K. D., Lozano, A. M., Tasker, R. R., and Dostrovsky, J. O., 1999. Pain-related neurons in the human cingulate cortex, Nat Neurosci, vol. 2, pp. 403-5.

[24] Adolphs, R., 2002. Neural systems for recognizing emotion, Curr Opin Neurobiol, vol. 12, pp. 169-77.

[25] Edwards, M. G., Humphreys, G. W., and Castiello, U., 2003. Motor facilitation following action observation: A behavioural study in prehensile action, Brain and Cognition, vol. 53, pp. 495-502.

[26] Kilner, J. M., Paulignan, Y., and Blakemore, S. J., 2003. An interference effect of observed biological movement on action, Current Biology, vol. 13, pp. 522-525.

[27] Chartrand, T. L. and Bargh, J. A., 1999. The chameleon effect: the perception-behavior link and social interaction, J Pers Soc Psychol, vol. 76, pp. 893-910.

[28] Decety, J. and Chaminade, T., 2003. Neural correlates of feeling sympathy, Neuropsychologia, vol. 41, pp. 127-138.

[29] Saygin, A. P., 2007. Superior Temporal and Premotor Brain Areas Necessary for Biological Motion Perception, Brain, vol. 130, pp. 2452-2461.

[30] Byrne, R. W., Barnard, P. J., Davidson, I., Janik, V. M., Mcgrew, W. C., Miklosi, A., and Wiessner, P.,

2004. Understanding culture across species, Trends Cogn Sci, vol. 8, pp. 341-6.

[31] Jacob, P. and Jeannerod, M., 2005. The motor theory of social cognition: a critique, Trends in Cognitive Sciences, vol. 9, pp. 21-25.

[32] Billard, A., Epars, Y., Calinon, S., Schaal, S., and Cheng, G., 2004. Discovering optimal imitation strategies, Robotics and Autonomous Systems, vol. 47, pp. 69-77.

[33] Decety, J., Chaminade, T., Grezes, J., and Meltzoff, A. N., 2002. A PET Exploration of the Neural Mechanisms Involved in Reciprocal Imitation, Neuroimage, vol. 15, pp. 265-72.

[34] Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., and Rizzolatti, G., 1999. Cortical mechanisms of human imitation, Science, vol. 286, pp. 2526-8.

[35] Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D., 2004. Empathy for pain involves the affective but not sensory components of pain, Science, vol. 303, pp. 1157-62.

[36] Chaminade, T., Meary, D., Orliaguet, J. P., and Decety, J., 2001. Is perceptual anticipation a motor simulation? A PET study, Neuroreport, vol. 12, pp. 3669-3674.

[37] Decety, J. and Chaminade, T., 2003. When the self represents the other: A new cognitive neuroscience view on psychological identification, Consciousness and Cognition, vol. 12, pp. 577-596.

[38] Oztop, E., Franklin, D., Chaminade, T., and Gordon, C., 2005. Human-humanoid interaction: is a humanoid robot perceived as a human, Internation Journal of Humanoid Robotics, vol. 2, pp. 537-559.

[39] Chaminade, T., Franklin, D., Oztop, E., and Cheng, G., "Motor interference between Humans and Humanoid Robots: Effect of Biological and Artifical Motion," presented at International Conference on Development and Learning, Osaka, 2005.

[40] Brass, M., Bekkering, H., Wohlschlager, A., and Prinz, W., 2000. Compatibility between observed and executed finger movements: comparing symbolic, spatial, and imitative cues, Brain Cogn, vol. 44, pp. 124-43.

[41] Gallagher, H., Jack, A., Roepstorff, A., and Frith, C., 2002. Imaging the intentional stance in a competitive game, Neuroimage, vol. 16, pp. 814.

[42] Press, C., Bird, G., Flach, R., and Heyes, C., 2005. Robotic movement elicits automatic imitation, Brain Res Cogn Brain Res, vol. 25, pp. 632-40.

[43] Press, C., Gillmeister, H., and Heyes, C., 2006. Bottom-up, not top-down, modulation of imitation by human and robotic models, Eur J Neurosci, vol. 24, pp. 2415-9.

[44] Gazzola, V., Rizzolatti, G., Wicker, B., and Keysers, C., 2007. The anthropomorphic brain: the mirror neuron system responds to human and robotic actions, Neuroimage, vol. 35, pp. 1674-84.

[45] Tai, Y. F., Scherfler, C., Brooks, D. J., Sawamoto, N., and Castiello, U., 2004. The human premotor cortex is 'mirror' only for biological actions, Curr Biol, vol. 14, pp. 117-20.

[46] Itoh, K., Miwa, H., Matsumoto, M., Zecca, M., Takanobu, H., Roccella, S., Carrozza, M. C., Dario, P., and Takanishi, A., "Various emotional expressions with emotion expression humanoid robot WE-4RII," presented at First IEEE Technical Exhibition Based Conference on Robotics and Automation (TExCRA '04), Tokyo, Japan, 2004.

The utility of gaze in spoken human-robot interaction

Maria Staudte
Department of Computational Linguistics
Saarland University
Saarbruecken, Germany
masta@coli.uni-saarland.de

Matthew Crocker
Department of Computational Linguistics
Saarland University
Saarbruecken, Germany
crocker@coli.uni-saarland.de

ABSTRACT

Psycholinguistic studies of situated language processing have revealed that gaze in the visual environment is tightly coupled with both spoken language comprehension and production. It has also been established that interlocutors monitor the gaze of their partners, so-called "joint attention", as a further means for facilitating mutual understanding. It is therefore plausible to hypothesise that human-robot spoken interaction would similarly benefit when the robot's language-related gaze behaviour is similar to that of people, potentially providing the user with valuable non-verbal information concerning the robot's intended meaning or the robot's successful understanding. In this paper we report preliminary findings from an eye-tracking experiment which investigated this hypothesis in the case of robot speech production. Human participants were eye-tracked while observing the robot and were instructed to determine the 'correctness' of the robot's statement about objects in view. Specifically, we examined the human behaviour in response to incongruency of the robot's gaze behaviour and/or errors in the statements' propositional truth. We found evidence for both (robot) utterance-mediated gaze in human-robot interaction (people look to the objects that the robot refers to linguistically) as well as for gaze-mediated joint attention, i.e. people look to objects that the robot looks at. Our results suggest that this kind of human-like robot-gaze is useful in spoken HRI and that humans react to robots in a manner typical of HHI.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; J.4 [**Social and Behavioral Science**]: Psychology

Keywords

gaze, joint attention, incongruency, utility

1. MOTIVATION

People have developed very subtle and complex strategies to communicate effectively, seamlessly integrating a variety of non-verbal signals during spoken language communication. Gaze as well as gestures, facial expressions and para-verbal feedback constitute some of these signals and they enrich communication in many social aspects and establish robustness. They help to convey information about attitude, emotional or belief state or simply coordinate the conversation by indicating turn-taking actions and let the partner know what the current focus of interest is. Psychological studies have revealed, for example, that gaze in the visual environment is tightly coupled with both spoken language comprehension [7, 8, 14] and production [10, 4]. It has also been established that interlocutors monitor the gaze of their partners (see e.g. [3] for a comprehensive account of joint attention). It is therefore plausible to hypothesise that human-robot spoken interaction would similarly benefit when the robot's language-related gaze behaviour is similar to that of people: not only would such behaviour imply human-like language processing, but it also provides the user with valuable non-verbal information concerning the robot's intended meaning (during robot production) or the robot's successful understanding of a user utterance (during robot speech recognition). In this paper we present work in progress and report findings from an eye-tracking experiment which investigated this hypothesis in the case of robot speech production.

Considerable work has already been done on gaze in HHI as well as robot gaze in HRI, e.g. during turn-taking [2] or with respect to information structure of the generated utterance [12]. Robot gaze generally in conversational engagement and in relation to some reference resolution has been explored by [13] among others and it could be established that the perception of robot gaze is coupled to the robot's head orientation [6]. The psychological findings from HHI, that have motivated our work, however, have not yet been applied in HRI . The role of utterance-mediated gaze in production as being tightly coupled to overall apprehension of an utterance has been established by [4], for instance. It has been shown, for example, that referential gaze is part of the planning process of an utterance and, thus, precedes the onset of the corresponding linguistic reference by approximately 800msec - 1sec. [9]. On the other hand, studies investigating gaze in comprehension, have revealed that listeners use speakers' gaze to identify a target before the linguistic point of disambiguation which clearly distinguishes utterance-mediated and gaze-mediated visual attention [5]. This study shows that gaze helps to identify possible referents of an utterance, even when the speaker's gaze was initially misleading due to the experimental setup. Subjects could establish a mapping of the speaker's gaze to their own visual scene and, thus, make use of the speaker's gaze during comprehension nevertheless. It is not clear that these insights from investigations of human cognitive behaviour can be mapped directly onto human-robot communication.

Robots differ in many ways as their physical means are distinct from ours. Robots do not possess the same amount of experience and world knowledge nor are they typically familiar with our communicative conventions. Hence, it is our general aim to investigate to what extent insights from human utterance-mediated gaze behaviour are sensibly applicable to robot gaze.

Our interest and the presented study focus on utterance-driven gaze behaviour by the robot, e.g. fixations towards an object before it is mentioned. Human gaze can then be observed in response to both the robot's speech (utterance-mediated attention) and the robot's gaze itself (joint attention). We conducted an initial experiment to show that our experimental design is generally valid and yields objective measures like decision/response times as well as the distribution of fixations to regions in the scene (which may bear evidence for other subjective/social factors). The scenario we have created is that of a robot describing a situation in blocksworld manner and simultaneously producing fixations to referenced objects. Human participants were eye-tracked while observing the robot and were instructed to determine the 'correctness' of the statement. Induced errors include incongruency of the gaze behaviour and/or errors in the statements' logical truth. These potentially reveal both the subject's attitude towards the robot as well as the utility of robot gaze in assessing validity of the robot's statements.

2. EXPERIMENT

2.1 Purpose and requirements

The presented pilot study aims to provide general empirical support for our hypothesis and method. Thus, its results provide only cues for further research trying to answer questions concerning the utility of robot gaze. If, indeed, gaze is a significant element of HRI then we can assume that *inappropriate* gaze behaviour may lead to some kind of disruption or slow-down in communication. In contrast, when the behaviour is consistent with (yet to be established) HRI conventions, we might expect interaction to be more fluent and efficient and, consequently, the acceptability and naturalness to rise. In this case, our longer term goal will be to find out what those gaze conventions are and what constitutes optimal robot gaze.

To begin investigating these issues, we require an experimental design that allows us to control the type and the occurrence of gaze and speech errors that might occur in robot speech production. Simultaneously, a method is desired that enables the experimenter to precisely observe the human subject and measure the reaction. A video-based setup fulfills these conditions by allowing the experimenter to very carefully plan and control errors and timing off-line while the subject's reaction can be recorded using an on-line eye-tracking technique. Although it might be argued that this is not real interaction, it has been shown that a video-based scenario without true interaction yields similar results to a live-scenario and can be considered to provide (almost) equally valuable insights into the subject's perception and opinion [17].

2.2 Methods

2.2.1 Participants

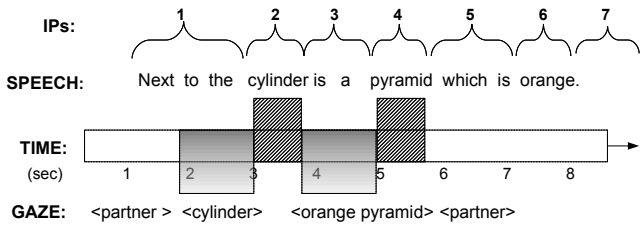
Ten students of various subjects, all enrolled at Saarland University and native speakers of German, took part in this pilot study. They had mostly no experience with robots nor with eye-tracking. They were told that the eye-tracker camera was monitoring their pupil size and, thus, the cognitive load of the task on them.

2.2.2 Material

Each video-clip showed a PeopleBot robot [1] onto which a stereo vision camera on a pan-tilt-unit was mounted, as it stood behind a table with a set of coloured objects in front of it. The objects were plain geometrical shapes of different colours. Two objects of the same shape - but of different colours - were target and distractor objects in a corresponding sentence. The video-clips each showed a sequence of camera-movements (that are called *fixations* for the human eye) towards either an object on the table or the assumed interaction partner, i.e. straight ahead. At the same time, a synthesised sentence of the following form was played back:

(1) a. "Next to the cylinder is a pyramid which is orange."
b. "Next to the <ANCHOR> is a <TARGET> which is <COLOUR>. *(as coded for analyses)*
c. "Neben dem Zylinder steht eine Pyramide die orange ist." *(original german sentence)*

The robot fixations and the spoken sentence were timed such that a fixation towards an object happened approximately one second prior to the onset of the referring noun which is consistent with psychological findings about the co-occurrence of gaze and referring expressions in human-human interaction [4, 16]. We can thus study two types of reactive human gaze: one being elicited by robot gaze (joint attention), the other being utterance-mediated (inspecting mentioned objects).

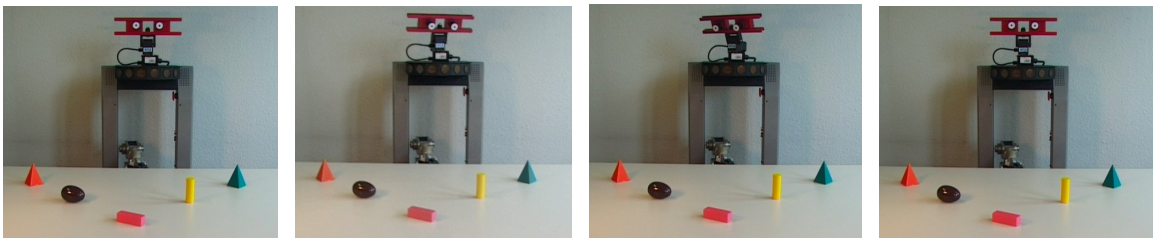


2: The timing of utterance-driven robot gaze, for sentence (1)

The presented videos were segmented into interest areas (IA) by means of bitmap templates, i.e. each video contained regions that were labelled e.g. "head", "table". Thus, the output of the eye-tracker could be mapped onto these templates yielding a certain number of hits for each IA. The spoken utterance is a sentence like example (1) describing the relation between a couple of objects. For our analysis the "cylinder" is encoded as the **anchor** reference and object, the "pyramid" is the **target** reference but may refer to the **target** object *or* the **distractor** object since there are two objects of the same shape, and the adjective "orange" is the linguistic point of disambiguation (LPoD). A similar design, also featuring late linguistic disambiguation with early visual disambiguation by means of gaze-following, was already successfully tested in a study on human-human interaction by [5].

Based on the onsets and offsets of the encoded linguistic events we segmented the video/speech stream into 7 interest periods (IP). The

[1] very kindly provided by the DFKI CoSy group: http://www.dfki.de/cosy/www/index.html and much appreciated



(a): Gaze straight ahead... (b): ...to anchor, (c): ...to distractor, (d): ...and straight ahead.

1: Frame sequence depicting a false-gaze condition (ti or fc) for sentence (1).

IPs encode the distinct time regions when the robot fixates an object and when it refers linguistically to an object (see figure 2). IP 7 is special, as it encodes the response time of the subjects, i.e. from the LPoD until the button pressing event, and is therefore a dependent variable. Because we are interested in the fixations occurring during that time period we included it as an interest period in our analyses. Note that although IP 7 varies in length it is typically longer than the other IPs and hence more fixations occur within it. This IP is to be analysed by itself with focus on the differences among the conditions.

Condition	**Spoken sentence:** / **Gaze towards:**
true - congruent (tc)	Next to the cylinder is a pyramid which is turquoise.
	<cylinder> <turquoise pyramid>
true - incongruent (ti)	Next to the cylinder is a pyramid which is turquoise.
	<cylinder> <orange pyramid>
true - no gaze (tn)	Next to the cylinder is a pyramid which is turquoise.
	<no gaze>
false - congruent (fc)	Next to the cylinder is a pyramid which is orange.
	<cylinder> <orange pyramid>
false - incongruent (fi)	Next to the cylinder is a pyramid which is orange.
	<cylinder> <turquoise pyramid>
false - no gaze (fn)	Next to the cylinder is a pyramid which is orange.
	<no gaze>

3: 3 x 2 conditions and samples.

A set of six items was constructed and each item was created in all six conditions resulting in a total of 36 video clips. The six conditions are shown in figure 3. We manipulated robot gaze behaviour as follows: gaze towards the correct target in the context of the described scene, gaze towards an incorrect object and no gaze during the utterance at all. Each gaze behaviour appears with a true or false statement about the spatial relation between two objects. The result is a set of six conditions: a true statement with no gaze (tn), with congruent correct gaze (tc) or with gaze towards an incorrect distractor object (ti), and a false statement combined with no gaze (fn), congruent gaze towards the mentioned but incorrect distractor object (fc) or incongruent gaze towards the correct but not mentioned object (fi). As shown in example (1) and figure 3 the general sequence of events in each item is as follows: the robot fixates the anchor object and then refers to the anchor linguistically, then - depending on the condition - the robot looks at the target or distractor object and subsequently it refers to either object linguistically before reaching the point of linguistic disambiguation (LPoD) which is the utterance of the colour towards the end of the sentence. The robot then looks back up towards the interlocutor. Note, that in the *no gaze*-conditions, the robot performs a quick glance at the visual scene before starting to speak and then remains still. This is to ensure that even though there is no relevant robot gaze behaviour the scene looks more or less natural.

2.2.3 Procedure

An EyeLink I head-mounted eye-tracker monitored participants' eye movements. The video clips were presented on a 21-inch color monitor. Viewing was binocular, although only the dominant eye was tracked and participants' head movements were unrestricted. For each trial, a video was played and its last frame remained on the screen until an overall duration of 11 seconds was reached. After a drift correction interlude the next video clip was presented. Prior to the experiment, the participants were instructed by a short text to attend to the scene and decide whether the robot was right or wrong. They were told that the results were used as feedback in a machine learning procedure for the robot. Next, the camera was setup and calibrated manually using a nine-point fixation stimulus. The entire experiment lasted approximately 25 min.

2.2.4 Predictions

If, indeed, this cognitively motivated robot gaze behaviour is beneficial, we expect incongruent gaze behaviour to cause a slow-down in cognitive processing measurable by recording decision/response times and possibly disruptions in fixations. Concerning response times we expect generally slower response times for false statements as this is a typical effect reported of in the literature (e.g. in the case of response times for match/mismatch tasks [15]). The three gaze conditions by themselves (*congruent, incongruent, no gaze*) are also expected to yield differences: congruent gaze should facilitate understanding and elicit faster response times than incongruent gaze. In the neutral *no gaze*-conditions there are two possibilities. Either this condition elicits the fastest response times because participants generally pay little attention to the robot's gaze, simplifying on-line information complexity. Or the neutral condition's response times lie between the *congruent* and *incongruent* conditions since there is neither supportive nor disruptive information conveyed.

With respect to participants' fixations we expect to observe gaze-following. That is, we predict that people fixate those objects or regions that the robot looks to. When the robot gazes towards the (incorrect) distractor object we still predict an increase in (gaze-mediated) looks towards the distractor by the participants. Generally, we anticipate that incongruent gaze behaviour - when robot gaze and robot utterance refer to distinct and incompatible objects - will elicit saccades between these two objects (target and distractor). Furthermore, we expect to observe utterance-mediated gaze.

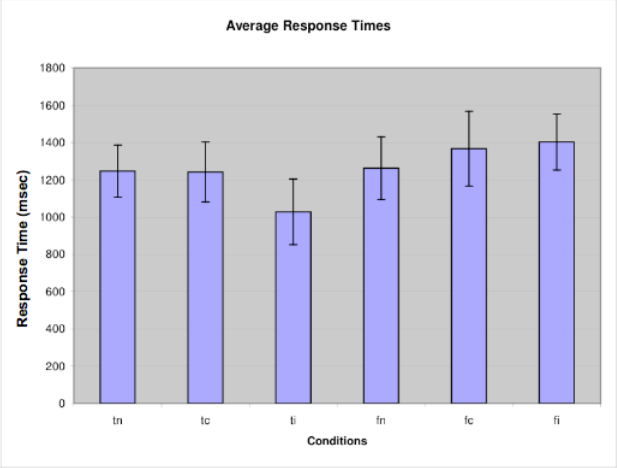
Once the robot's speech identifies an object or scene region we predict increased looks by our participants towards this region.

The most critical IPs, with regard to our predictions on gaze-mediated looks, are IP 1 and 3, which correspond to robot gaze movements. Most critical, with respect to utterance-mediated fixations, are IPs 2 and 4 (and possibly the subsequent one respectively) which correspond to the periods when the robot refers to an object linguistically. Further we similarly anticipate saccades between the target and the distractor during the response time period (IP 7) because we expect people to visually assure their decision and check all possible referents before giving the answer.

3. RESULTS

Response Times

The analysis of the response times, i.e. the time from the mentioning of the LPoD until a button press for either 'true' or 'false' was recorded, revealed that the *false-incongruent* condition (fi) results in the slowest response (figure 4). An unpaired t-test confirmed that *fi*-responses were significantly slower than *ti*-responses (*true-incongruent*): the difference in means is 375.13 msec with a 95% confidence interval of $375.13 \pm 1.96 * 117.85 = (144.14, 606.1)$ with t(*ti*,*fi*)$= 3.18 > t(p < 0.001, df = 128)$.



4: Average response times in msec for each condition, including upper and lower bounds for

The expected and observed general tendency for wrong statements to elicit longer response times than true statements is apparent in the graph as well. The *no gaze*-conditions are neither faster nor slower than the gaze-conditions which suggests that people do make use of robot gaze and are not finding it distracting or annoying (even though it often is wrong in this study). The slow response time for *false-incongruent* trials suggests that the participants had difficulty to determine correctness especially when a statement was false (i.e. the robot referred to the wrong object) although the robot was fixating an object that would have been correct to mention in this situation. This is consistent with our hypothesis that robot gaze is useful. In particular when it is used correctly by the robot, the gaze modality becomes a competitor to the language modality - at least in those cases where the utterance conveys unexpected or wrong information.

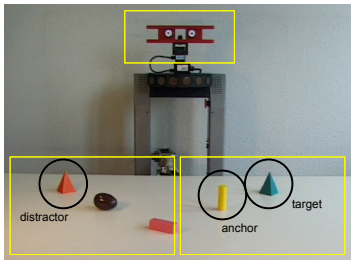
Furthermore the condition *true-incongruent* yields considerably faster response times than the other two *true* conditions (tc, tn). This initially surprising result still supports the hypothesis that gaze is useful - even when it is wrong - by yielding faster results than the *no gaze*-conditions. Considering the design of the pilot study, i.e. without fillers and the same gaze and sentence pattern for each trial, it is not surprising that people adjust and learn to recognise wrong gaze behaviour faster. In fact, this may cause some distortion of the response times in general. However, when both robot gaze and the spoken sentence are congruently referring to a wrong object (i.e. the statement is logically false), the response times are still considerably slower than the remaining four conditions (tn, fn, tc, ti). That suggests that even though both modalities are wrong, and obviously so, their congruency elicits longer response times and, hence, seems to pose a higher cognitive load.

Another interesting effect is revealed by the number of incorrect answers and those that were not given at all. It occurred several times that subjects did not press a button at all. Out of 8 omitted answers, 6 occurred in a *true-no gaze* condition and 2 in a *true-incongruent* condition. Incorrect answers were given in 22 trials, out of which 14 occurred in an *incongruent* condition and 5 in a *no gaze* condition. This makes an overall error of 7 % of all trials. The omitted and incorrect answers in trials featuring *incongruency* add up to 16 (53 % of all errors) and in trials without any directed robot gaze there are 11 incorrect answers (37 % of all errors), whereas only 3 incorrect answers where found in *congruent* trials. Again, this supports the claim that (congruent) gaze contributes to successful understanding even when produced by a robot.

Fixations

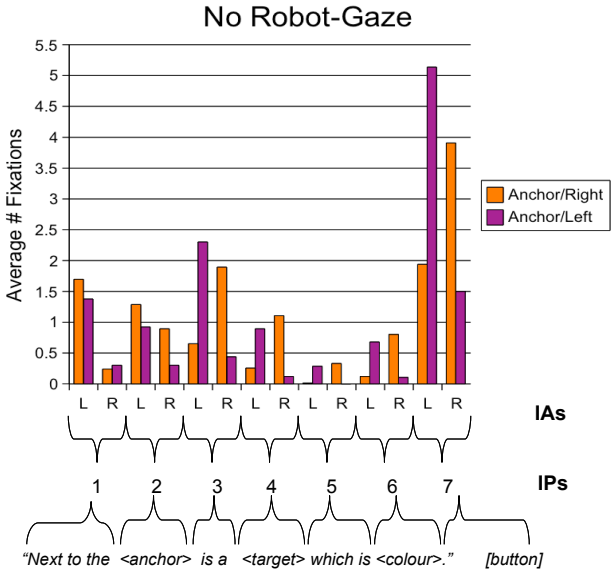
An initial analysis of the average number of fixations over all subjects per condition and per interest period (IP) and interest area (IA), here robot head and table, shows that there is a general rise in absolute number of fixations to the table area as soon as the first object is mentioned. After a slight decline towards the end of the sentence the number rises again considerably in the time period between the LPoD and the moment the subject presses the button. During the same IP the average number of fixations on the robot head rises as well. This may be due to the relief of concentration after the sentence has ended (and people had time to inspect the head) or may simply be the default gaze direction, i.e. straight ahead.

In a more fine-grained analysis, we have looked at three IAs, one being again the robot head and two more where the table area has been divided into two parts, left and right. In half of the trials the anchor object and the target object are positioned in the right half of the table area whereas the distractor object lies in the left area of the table, and vice versa for the other half. We therefore refer to the area that contains the referent and target objects as the target area and to the other side of the table as the distractor area. This kind of segmentation allowed us to observe whether participants followed the robot's gaze movement without fixating the



5: IAs 'head', 'left' and 'right'

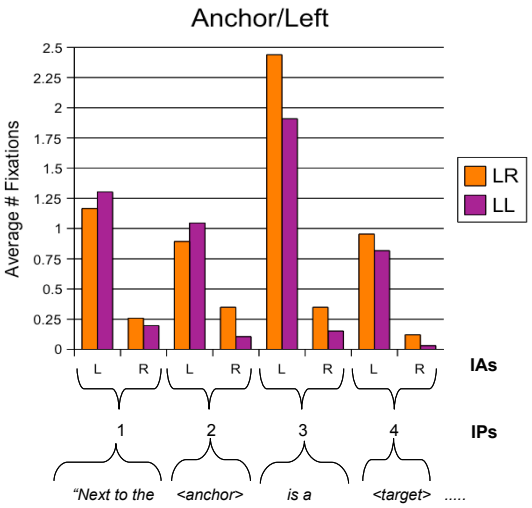
robot head directly. As a result we observed a general bias for fixations on the left side of the table. This becomes evident in figure 6 which shows the control condition *no gaze* (tn, fn). IP 1 and 2 in this chart show a clear preference for looks towards the left side of the table independently of where the anchor/target objects are positioned. This bias is commonly observed, reflecting general human scan patterns. From IP 3 on, however, utterance-driven fixations can be observed showing the expected preference for the area containing the mentioned anchor/target.



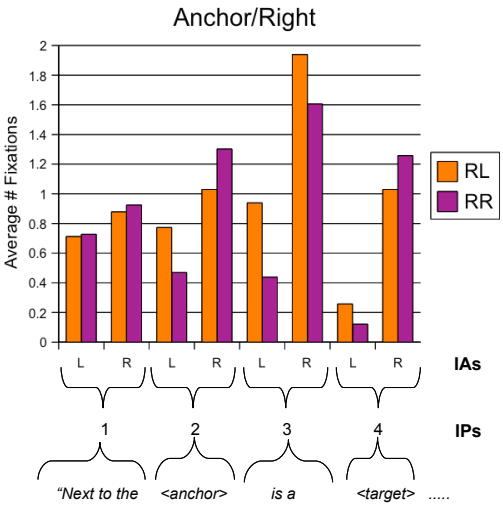
6: Average number of fixations in the neutral *no gaze*-conditions, with target being on either the right or left side. L and R on the x-axis denote the IAs 'left' and 'right'.

In figure 7 we have plotted the average number of fixations in IPs 1-4: 1, 3 depicting mainly robot gaze-mediated fixations (joint-attention) and IP 2, 4 showing mainly utterance-mediated fixations. Note, that RR, RL etc. denote the direction and therefore congruency of gaze movements, i.e. towards the right side of the table (target area) and again rightwards is correct gaze movement, the abbreviation is thus RR. RL indicates gaze movement towards the anchor (target area) and then to the left side of the table where the distractor object is located. For IPs 1 and 3 we detected more fixations on the anchor/target area than in the distractor area, notably already before the object was actually mentioned. The result is significant according to the paired t-test, for instance in IP 3 for the target on the left side: mean difference $\bar{x}_L = 1.92$ fixations in a confidence interval (1.68, 2.15) with $t(\bar{x}_L) = 16.13 > t(p < 0.001, df = 131)$. And accordingly for the target on the right side: mean difference $x_R = 1.08$ fixations in a confidence interval (0.805, 1.354) with $t(\bar{x}_R) = 7.714 > t(p < 0.001, df = 131)$.

The same effect, i.e. a significant rise in fixations towards the object that is (now linguistically) referred to, is visible in IPs 2 and 4 and is continued throughout the rest of the trial. The most critical region is IP 3 where the robot's gaze is either turned towards the target or the distractor object at the other side of the table. At this stage it becomes evident whether subjects believe that the robot gaze is an early indicator of what is going to be mentioned next. Our recordings reveal only a slight increase of fixations towards the distractor as a reaction to robot gaze towards the distractor object. The experimental design may prevent a stronger effect because the repeated gaze pattern in the items allows the participants to predict what is going to happen.



(a): Here IA "L" contains anchor and target...

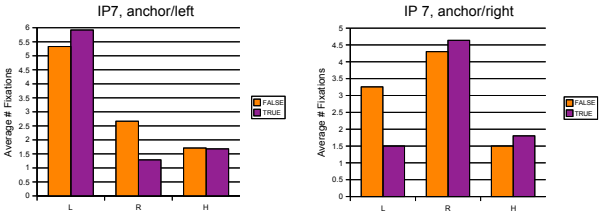


(b): ...while here IA "R" contains anchor and target.

7: Anchor/target on one side of the table, distractor on the opposite side. Depicted are IPs 1-4 showing gaze- and utterance-mediated fixations.

IP 7 is plotted in figure 8 which depicts the distinct conditions within the decision time period, i.e. from the disambiguating adjective until the button press event. This reveals that a false statement leads to a significantly higher number of fixations in the distractor area than is the case for true statement trials: mean difference of average fixations on the distractor area and on target area (for target/left side) is $\bar{x}_L = 1.38$ fixations in a 95% confidence interval (0.645, 2.115) with (un-paired) $t(\bar{x}_L) = 3.373 > t(p < 0.001, df = 130)$, and similarly for target/right side: $\bar{x}_R = 1.755$ fixations in a confidence interval (1.003, 2.507) with $t(\bar{x}_R) = 4.5726 > t(p < 0.001, df = 130)$. This result is not surprising as giving wrong answers typically affords slightly longer response times (as reported in the previous section) which should also be reflected in the direction and number of fixations performed during

that time. However, we did not observe the expected difference with respect to congruency at this stage. That is, a true statement yielded similar fixation results independently of the correctness of the robot-gaze. It is likely that this is due to the relatively easy spatial arrangement and the long time period until a decision is demanded such that participants can look around extensively beforehand.



(a): Distractor object lies on the right...

(b): ...and here on the left side.

8: Fixations for true/false statements during response time period.

4. DISCUSSION

We found clear evidence for (robot) utterance-mediated gaze in human-robot interaction: people look to the objects that the robot refers to linguistically. This is not a surprising result but it is useful nonetheless as it confirms typical human behaviour in response to robot speech and gaze (even in video-based interaction). Further evidence was collected during the response period: we registered a strong tendency of the participants to fixate both the target and the distractor object when the statement was false. In those cases, the uttered sentence referred to the distractor object linguistically and people looked more often towards the distractor object than in those trials where a true sentence was uttered. We also found clear evidence for gaze-mediated joint attention, i.e. people look to objects that the robot looks at. IP 1 was the period immediately preceding the linguistic reference to the anchor and already then participants looked towards the anchor. These results support our hypothesis that human-robot spoken interaction is governed by principles similar to HHI. When the robot's language-related gaze behaviour is similar to that of people we observe human gaze patterns that are typical for HHI.

Moreover, the reported response accuracies suggest that generally incongruent robot behaviour (i.e. divergence of both modalities speech and gaze) is causing confusion. The measured response times are slightly more difficult to interpret. False statements elicited slightly longer response times which is typical human behaviour. We also found that the *false-congruent* condition response times were significantly slower than in the *true-incongruent* condition (which seems to contradict the evidence from response accuracy). However, this could suggest that the coherence of the modalities in the (wrong) fc-condition leads to stronger doubts about the truth of the statement than in the ti-condition. The ti-condition in which robot gaze is wrong while the linguistic statement is true seems to allow fast reference resolution. Considering the design of the pilot study, i.e. without fillers and the same repetitive gaze and sentence pattern in each trial, we assume that people adapt to the task and learn to recognise early when gaze is erroneous which may generally distort response times.

The effects we found were not always in accordance with our predictions. We assumed, for instance, that incongruent robot-behaviour elicits more fixations on both potential referents, target and distractor, during the linguistic utterance and during decision making. This was partially observed in the latter interest period for false statements. The direction of the robot gaze, however, seemed to be irrelevant for the final decision process. For wrong gaze we did not observe a particular rise in fixations towards the distractor area in IP 3 either. This IP comprises the robot-gaze movement towards the target or distractor object and, thus, reports gaze-following. Presumably this again is due to the fact that the course of events in a trial becomes predictable after a while.

5. CONCLUSIONS AND FUTURE WORK

We have shown that, in principle, it is possible to use detailed insights into human cognition and behaviour to enrich human-robot-interaction. The presented evidence shows that this kind of robot-gaze is beneficial in HRI and that humans react in a manner typical of HHI to both robot speech and robot gaze. We predicted that in case one or both robot modalities are infelicitous a slow-down of the interaction would be measurable by response times and fixation distributions. Our results support this hypothesis and reveal several cases where incongruent robot behaviour leads to slower response times or disruptions in the usual distribution of fixations.

The presented study also shows that the methods we used to measure the effects and success of the robot behaviour objectively during robot production are generally appropriate and effective. However, we found some weaknesses in the experimental design such that the obtained results, although promising, are only preliminary. The spatial arrangement of the scene, for instance, is small and simple. A larger area with more complexity could lead to clearer results concerning the robot gaze-mediated fixations, i.e. the robot gaze could be considered more useful for early referent resolution. The presentation of the items is going to be interleaved with the presentation of filler videos which differ from the items and enforce gaze reliability. This ensures that the participants will not be able to predict what the robot is going to do. More crucially, the presentation of mostly *true-congruent* fillers will influence the trade-off between cost and benefit of robot-gaze for information processing during communication: the more errors occur in the trials the likelier will participants decide to ignore gaze as a source of information. If, however, gaze is mostly a useful early indicator for correct reference resolution, its benefit for on-line comprehension should override the extra costs caused by errors. Furthermore, we have put a lot of emphasis on congruency in this study but designed a task for the subjects that focusses on the truth value of the linguistic statement alone. In order to emphasise the effect of incongruency we plan to change the task such that subjects are required to consider the robot performance more holistically, e.g. "If you think the robot is wrong, tell it what is wrong". There is a trade-off here between simplicity in measurements: we lose the option to record response times but we gain additional information on what the participants expected of the robot and what they actually perceived. This is possible because incongruency arises from different errors made by the robot: either the linguistic reference is incorrect or the gaze is incorrect. This failure may be attributed to incorrect wording (as in 'correct gaze but false statement' = ti), incorrect judgement of the spatial relations, errors in visual (colour) processing, erroneous gaze movement etc. Thus, by asking the participants to actively decide which object was meant and why the error occurred, we also learn more about the "theory-of-mind" (ToM) that people ascribe to the robot and how robot gaze contributes to forming a "theory-of-mind" (see e.g. [1] for work on ToM among humans and [11] for

research on apes). With increasing communicative competence of robots it becomes more and more interesting to investigate people's attitude towards robots.

6. ACKNOWLEDGMENTS

The research reported of in this paper was supported by IRTG 715 "Language Technology and Cognitive Systems" funded by the German Research Foundation (DFG).

7. REFERENCES

[1] S. Baron-Cohen, A. Leslie, and U. Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21:37Ð46, 1985.

[2] J. Cassell, O. Torres, and S. Prevost. Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. *Machine Conversations*, pages 143–154, 1999.

[3] P. D. Chris Moore, Philip J. Dunham, editor. *Joint Attention Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[4] Z. M. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11:274–279, 2000.

[5] J. Hanna and S. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57:596–615, 2007.

[6] M. Imai, T. Kanda, T. Ono, H. Ishiguro, and K. Mase. Robot mediated round table: Analysis of the effect of robot's gaze. In *Proceedings of 11th IEEE ROMAN '02*, pages 411–416, 2002.

[7] P. Knoeferle and M. Crocker. The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, 30:481–529, 2006.

[8] P. Knoeferle and M. Crocker. The influence of recent scene events on spoken comprehension: evidence from eye-movements. *Journal of Memory and Language (Special issue: Language-Vision Interaction)*, 57:519–543, 2007.

[9] A. Meyer, A. Sleiderink, and W. Levelt. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66:B25–B33, 1998.

[10] A. Meyer, F. van der Meulen, and A. Brooks. Eye movements during speech planning: Talking about present and remembered objects. *Visual Cognition*, 11:553–576, 2004.

[11] B. H. Michael Tomasello, Josep Call. Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences*, 7:1632–1634, 239-240.

[12] B. Mutlu, J. Hodgins, and J. Forlizzi. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In *Proceedings 2006 IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS'06)*, Genova, Italy, 2006.

[13] C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[14] M. K. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, 1995.

[15] G. Underwood, L. Jebbett, and K. Roberts. Inspecting pictures for information to verify a sentence: eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology*, 56:165–182, 2004.

[16] F. F. van der Meulen, A. S. Meyer, and W. J. M. Levelt. Eye movements during the production of nouns and pronouns. *Memory & Cognition*, 29(3):512–521, 2001.

[17] S. Woods, M. Walters, K. L. Koay, and K. Dautenhahn. Comparing Human Robot Interaction Scenarios Using Live and Video Based Methods: Towards a Novel Methodological Approach. In *Proc. AMC'06, The 9th International Workshop on Advanced Motion Control*, 2006.

A Visual Method for Robot Proxemics Measurements

Tim van Oosterhout
Instituut voor Informatica,
Universiteit van Amsterdam
Kruislaan 403 1098SJ
Amsterdam, The Netherlands
tjmooste@science.uva.nl

Arnoud Visser*
Instituut voor Informatica, Universiteit van Amsterdam
Kruislaan 403 1098SJ
Amsterdam, The Netherlands
arnoud@science.uva.nl

ABSTRACT

Human interaction knows many non-verbal aspects. The use of space, among others, is guided by social rules. Not conforming to these rules may cause discomfort or even miscommunication. If robots are to interact with people, they must follow similar rules. The current work tries to identify factors that influence human preferred interaction distance in conversation-like interaction.

For the measurement of interaction distances an accurate and objective visual method is presented. In this method, the researcher does influence the results by disturbing the interaction.

It is found that subjects choose interaction distances comparable to those in human interaction. Variations are mostly explained by subject age and, depending on age, by gender or robot appearance. This is the first time, to our knowledge, that a clear age and gender effect is found in human-robot interaction-distance.

Categories and Subject Descriptors

I.4 [**Computer Applications**]: Social and behavioral sciences — *Psychology*; I.2.9 [**Artificial Intelligence**]: Robotics — *Commercial robots and applications*; H.5 [**Information Systems**]: Information Interfaces and Presentation (e.g., HCI) — *Benchmarking*

General Terms

Measurement, Experimentation, Human Factors

Keywords

Proxemics, Human-Robot Interaction, Visual Measurement

1. INTRODUCTION

*The authors were supported by EU Integrated Project COGNIRON ("The Cognitive Companion") FP6-002020.

Advancements in artificial intelligence enable the creation of more intelligent robots that can perform a greater array of tasks, making it more realistic and even desirable to bring them into the house or office. People are social beings however, and human interaction is guided by social rules. While people can learn to adapt to robots, the robots should be made to follow similar rules that will make the interaction natural and require no extra effort on the human part. The current research tries takes an approach from the sociological concept of proxemics.

Recent research has indicated the influence of robot appearance [4], subjects' personality [8] and type of interaction on the interaction distance. The typical setups of those experiments were inside the laboratories, where colleagues/volunteers got clear assignments about the type of interaction that should be started. The research reported here is performed in a free setting during an arts and technology festival, with the subjects unaware of the experiment performed. This resulted in a large number of interactions, with variety in age and gender which is difficult to reproduce in robotics laboratories.

1.1 Proxemics

The field of proxemics is concerned with interpersonal distance and personal space. The term was coined by the anthropologist Edward T. Hall in his 1966 book *the hidden dimension*. In this book, Hall uses findings from the animal kingdom and insights in human experience of space to define four personal spheres. These spheres define areas of physical distance that correlate reliably with how much people have in common (cultural difference). Where the boundaries of these spheres exactly lie is additionally determined by factors such as gender, age and culture [2, 3, 5]. When one comes too close to another, the other may feel crowded or intimidated. If, on the other hand, one stays too far back, this is seen as awkward and one may be perceived as cold or distant. Appropriate distances found by Hall in western culture for adults of both genders are displayed in Table 1.

1.2 Human Interaction

To explain the locations of these boundaries, Hall theorizes that they coincide with the boundaries of sensory shift. At different distances, touch, vision, hearing but also smell may be optimal, distorted, or not available at all. Physical properties also come into play, such as an arm's length, which defines the distance from where one can touch the other, or two arms' length, which defines the boundary where interaction partners can cooperate to make physical contact [2].

Table 1: The four spheres of physical distance corresponding to cultural difference according to Hall.

Designation	Specification	Reserved for ...
Intimate distance	0 - 45 cm	Embracing, touching, whispering
Personal distance	45 - 120 cm	Friends
Social distance	1.2 - 3.6 m	Acquaintances and strangers
Public distance	> 3.6 m	Public speaking

1.3 Human-Robot Interaction

In proxemics studies, the focus lies on human-human interaction. However, when one interaction partner is a robot, it is not well known to what extent the different factors of human proxemics still apply and what new factors play a role. Moreover, since robots typically do not have an odor or body heat, sensory input can no longer explain or predict appropriate distances, even if the limitations on vision and hearing may still apply.

While human-robot proxemics may follow a similar pattern as human and animal proxemics in having distinct zones, no assumptions about such existence or the locations of possible boundaries are made in the current research. Instead, the focus is to identify factors that influence interaction distance and their effect. In Section 1.4, a list is presented of such possible factors, all of which were included in the empirical study. Along with the description of each factor, a rationale to include it is given. Factors that were not included were factors that are irrelevant for a robotic interaction partner, such as body heat or smell.

1.4 Included Factors

Robot type could count towards the cultural difference equivalent of human-robot interaction. People may prefer to interact with a robot with which they have more in common or with which interaction is easier due to the height and shape. This would translate into more frequent observations of interaction with a certain robot, but may also influence the preferred distance. Specifically robot height and shape was investigated.

Although Hall doesn't mention **subject height** as a factor, there are studies that do take it into account because height difference influences face-to-face distance [5]. In addition, when adjusting a screen or monitor, appropriate height and orientation are meant to achieve a neutral neck position and minimal neck movement at the optimal viewing distance. Since subjects had no control over screen height and orientation, they might have chosen a different distance instead to view the screen at a more comfortable angle.

Since **subject gender** is an important factor in human proxemics [3, 5], it may also play a role in human-robot proxemics. This point is complicated by the fact that the robots used in this experiment represented a person whose gender might be of influence. The operator's gender was left out of consideration however, since the operator's gender was only obvious for the Mobi Sr. robot, and its operator could change at any time (see Sections 2 and 3.1). Since the measurements are pooled, any gender effects found would then represent how men's and women's preference are different in regard to a genderless robot.

Subject age is a factor in human cultural difference and therefore in human proxemics [2, 1], thus it might also be of influence in human-robot proxemics. The same complication as with the operator's gender arises, and it is disregarded on the same grounds.

When the location of interaction is **crowded** with people, it may be impossible for a subject to keep the preferred distance since doing so might bring him or her undesirably close to one or more other people. Since only the upper bound of distance options is limited, subjects are forced to stand closer to the robot. However, in such a situation the subject is also forced to stand closer to other humans. It would be interesting to see how the subject resolves this shortage with respect to the relative amount of distance the subject gives up to the robot and other humans.

1.5 Hypotheses

Based on the inclusion rationales for each factor, we formed the following hypotheses:

- Children prefer the smaller robot, which means more observations with it and smaller distance compared to Mobi Sr.
- Height difference between subject and robot causes greater distance.
- Men will stand closer because of affinity for technology
- Younger people will stand closer as they do in human interaction [1].
- Spatial constraints caused by crowding cause smaller distance.

2. MATERIALS

Two robots were used in the current experiment. They could be controlled by volunteers through a desktop computer to which the robots were connected via a wireless network.

2.1 Robot 1: Mobi Sr

The first robot, called Mobi[1] Senior (Figure 1), was approximately 175 cm tall and had a round base with a diameter of 66 cm with semi spheres sticking out to cover the support wheels. It was driven by two wheels left and right of the centre of the base, and balanced by 4 passive wheels around the base. The robot was not made to resemble human form. In spite of this, it was intended to be a communication device. It was equipped with a monitor which was mounted at the top of the robot at eye level. This monitor showed a video feed that was sent from a webcam at the operator's computer showing the operator's head and shoulders, as is typical for a web conference. The robot had a webcam mounted directly above the robot's monitor allowing the operator to view

[1]Mobile Operated Bi-directional Interface

Figure 1: Two people interacting with Mobi Sr.

the remote location. In addition, the robot had stereo speakers and a microphone, and the operator's computer had a stereo headset and a microphone as well, enabling two-way audio communication between the operator and an interaction partner. The operator could move the robot back and forth and rotate it left or right around its axis by using the arrow keys on the local keyboard.

2.2 Robot 2: Mobi Jr

A second, smaller robot was used called Mobi Junior (Figure 2). It had many of the same features as Mobi Senior, with the most notable exceptions of lacking a monitor to show the operator, and being only 112 cm tall. It had a square base with rounded edges and was 60 cm wide and deep. In addition, Mobi Junior's operator also had camera controls to aim the camera anywhere between 28 deg up and 25 deg down. Its shape was quite different, but apart from this, Mobi Junior also had stereo speakers and a microphone, allowing the same two way audio communication, and a webcam mounted in a round head to allow the operator to see the remote location. Mobi Junior was designed to appeal to children, who might have trouble seeing the monitor on Mobi Senior and who would be too short to be seen by its camera, or who might be intimidated by such a tall mobile device.

3. METHODS

3.1 Setting

The robots were showcased during a three day arts and technology festival. This festival was held in a former factory and covered three large halls. There was a stand belonging to the Mobi team where visitors could volunteer to operate either robot. There were two large screens at the stand facing the hall where visitors could see the video feeds from the robot cameras. Both robots could be directed to any location within the halls from the stand through a local wireless network. Visitors were free to take control of the robots or to interact with them. People from the Mobi team were present at the stand to give information about the robots and instructions on how to

Figure 2: A girl interacting with Mobi Jr. and two bystanders (faces have been blurred).

control them, and at the robots' locations to answer any questions.

3.2 Procedure

To determine the appropriate distance the robots would need to keep, measurements were made on the distance to the robots that people voluntarily chose in different situations. A prospective observational design was chosen to ensure ecological validity. All approaches were voluntary and without knowledge of the experiment. Volunteer operators were not instructed to stop moving the robot during interaction with people, but consistently did so. Interactions were not included until the interaction was established and the robot had stopped. Digital photographs were taken of interactions and were analyzed later (see Section 4.4). Subjects were included more than once only if they were observed in different situations with respect to crowdedness, and only once per crowdedness category (see below). Photographs typically showed several people each, sometimes in a small crowd. It was not unusual to have more than one interaction per photograph, with a maximum of five. Out of all taken photographs, 72 were used for distance measurement, depicting 106 subjects in 140 observations.

Additionally, frequencies were collected of observed interactions between age group and robot type. Subjects were included only once, even if they were included more than once in distance measurement. Photographs that were unsuitable for distance measurement could be included in this tally if the pictured subject was not yet seen in the distance measuring photographs and if the interaction met the previously stated requirements. For age group/robot type frequencies, 135 unique subjects were counted.

Because of the observational nature of the experiment, subjects were not approached by the researcher to fill out any questionnaires. Therefore, subject length had to be measured on the photograph (see Section 4.4) and subject age was estimated. Because of the imprecise nature of estimation, age was restricted to four categories shown in Table 2.

Table 2: Age categories used in the proxemics experiment

Category	Ages	Notes
Children	0 - 11	Subjects predate puberty
Teenagers	11 - 19	
Young Adults	19 - 30	Subjects are typically students
Adults	> 30	

4. METRICS

4.1 Interaction

Observations were included if a subject directly interacted with the operator or the robot. For Mobi Sr. this could include talking with and waving or gesturing at each other. Even though subjects could not see the operator on Mobi Jr., waving at or touching the robot was also considered direct interaction. Observations were also included if another person interacted directly with the robot while the subject stood in front of the robot and faced it in a way that the subject too could interact with it, either through conversation or gesturing.

4.2 Crowdedness

Crowdedness was quantified with Hall's four spheres of personal distance in mind. It is determined as the biggest sphere in which the subject may choose to stand while still being able to interact with the robot. The actual distance the subject chooses can be classified at most as this sphere, or a smaller one. This is a per-subject classification which means that subjects in the same photograph may be assigned to different crowdedness categories.

Interaction is blocked if a person obscures the view between subject and robot. In this case, the interaction is not counted and no crowdedness category is assigned. It is possible for several people to directly interact with the robot if they stand next to each other. In this case the interaction would cease to be direct for one subject if the subject would move away, so to maintain the same type of interaction, the subject can at most be in the sphere he or she is already in. In that case the subject's distance directly determines the crowdedness category.

If a subject already interacts indirectly because another person stands closer without blocking the subject's interaction, then the interaction is counted and the crowdedness category is also assigned to the actually occupied sphere.

4.3 Measured Distance

The exact measured distance is usually nose-to-nose distances [1, 9, 3]. However, since in the current experiment one interaction partner lacks a nose, another measure had to be devised. Moreover, given the utilized measurement methods, accurate measurements could only be made for distances on a given plane, more specifically the floor. For these reasons, the point where the subject stood was defined as the point on the floor directly under the centre of the subject's torso. This point is a fair indication (though not an average) of the position of either foot and also takes leaning forward or backward into account. The measured distance was from this point to the nearest point on the robot's shell. For the robots, no central point was defined because their shells created a perimeter that could not be crossed, thereby defining a suitable minimum distance. Human beings on the other hand can stand over smaller objects, which can be expressed in the chosen scheme. In addition, neither robot could lean, so no corrections would have to be applied to the perimeter.

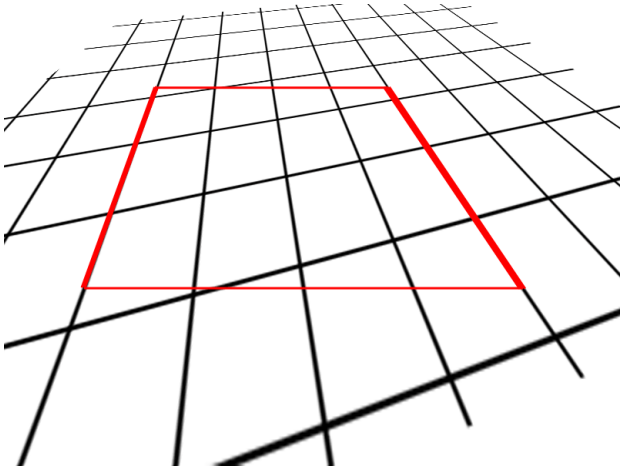


Figure 3: Parallel lines in perspective with a highlighted trapezoid constructed from a random pair of horizontal lines and the pictured perspective lines.

Note that in this scheme the measured distance is greater than 0 if the subject's feet are physically touching the robot's shell. A measured distance of 0 means that the subject has placed one foot on either side of the robot's base and is standing over it, which was theoretically possible with both robots, but only feasible with Mobi Jr. This measuring scheme gives measurements that are comparable to nose-to-nose distance for the Mobi robots. The contribution to the distance for a person standing upright will typically be almost a foot's length too long, but the robots contribution will be too short because their heads (the round head containing the camera for Mobi Jr., and the monitor for Mobi Sr.) are receded with respect to the base, and so would have given bigger measurements if measured from where their noses might have been if they had them.

4.4 Visual Measurement

Digital photographs were used to determine the distance to the robot chosen by the subject, subject height and the distance between the robot and the nearest person relevant to determine the crowdedness category. All photographs pictured the entire robot and the entire subject. If possible, the photograph was taken from a position perpendicular to the line between subject and robot.

In photographs where the subject and the robot were in a plane parallel to the camera's focal plane, perspective distortion was not an issue and distance measurement was very similar to the method used in [3]. Since all the measurements from the robot were known, a ratio between pixels and centimeters could easily be established. This ratio then related pictured lengths to actual lengths, with which subject distance and height could be measured. At the resolution the photographs were taken, robot height measurements ranged from about 600 to 2700 pixels,

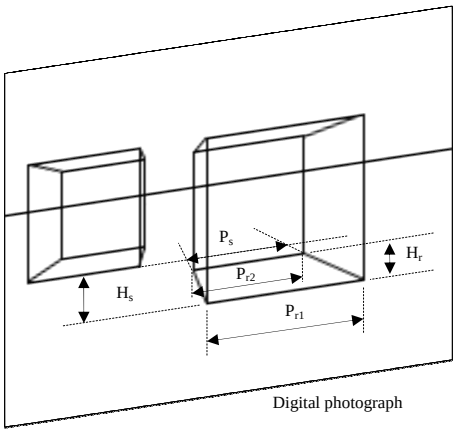


Figure 4: The reference object and the subject in a photograph.

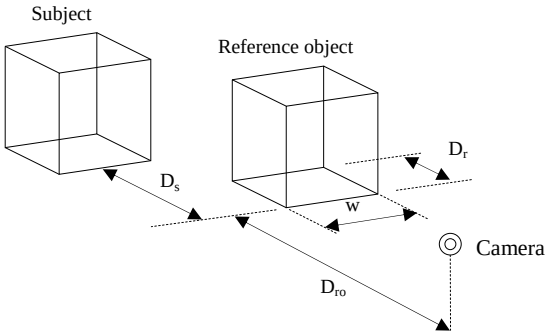


Figure 5: The reference object and the subject in the world.

but would typically be around 1600 pixels, giving subcentimeter precision for size measurements and distance measurements without a perspective element.

Even though no markings were applied, the floors in the former factory halls had enough features to find a pair of parallel lines. Another pair could be freely chosen in the picture out of any pair of perfectly horizontal lines, since these are always projected parallel to the camera's focal plane and thus to each other[2]. The two pairs will enclose a trapezoid in perspective projection (Figure 3). The ratio between the length of the top and bottom of the trapezoid, P_{r2} and P_{r1} respectively (Figure 4) provides the amount of decrease in size due to perspective distortion over a distance D_r (Figure 5) whose projected size is given by the height of the trapezoid H_r. This ratio may also be viewed as a scale factor, giving the size of objects projected on the top line P_{r2} in relation to objects projected on the bottom line P_{r1} or vice versa, provided that they reside on the same plane, such as the floor. Using the parallel lines that follow the reference plain (the floor), such a scale ratio can be calculated for any given height, for instance H_s, in the photograph by choosing another horizontal line P_s to form the top of the trapezoid. In this way, sizes of objects on the floor can be related to one another. By relating a position to that of the robot with known dimensions, sizes such as subject height can now be measured across the entire photograph.

To obtain the distance between any two points on a plane, a known reference distance D_r is needed. This reference distance serves to quantify perspective distortion and to relate projections with a depth component to actual size. This distance would need to be perpendicular to the focal plane. If the reference object is not aligned in such a manner, then a bounding trapezoid (projection of a rectangle) can be constructed with known measurements that is aligned in this way using the image centre and Pythagoras' theorem. Let us assume furthermore that the optical axis is parallel to the floor. Now take the distance D_{r0} from the camera to the reference object. Using the object's known width w, the projection of this width on the picture plane P_{r1} and the focal distance f (the distance between the focal point and the picture plane), we could directly compute the depth distance:

$$D_{r0} = \frac{w}{P_{r1}} f \tag{1}$$

A depth distance between two points in the photograph can be expressed as a difference between two absolute distances, e.g.:

$$D_r = \frac{w}{P_{r2}} f - \frac{w}{P_{r1}} f \tag{2}$$

When expressed as a ratio of distances, the focal distance f and reference measure w are eliminated:

$$\begin{aligned} \frac{D_r}{D_s} &= \frac{\frac{w}{P_{r2}} f - \frac{w}{P_{r1}} f}{\frac{w}{P_s} f - \frac{w}{P_{r1}} f} = \frac{\frac{1}{P_{r2}} - \frac{1}{P_{r1}}}{\frac{1}{P_s} - \frac{1}{P_{r1}}} = \frac{\frac{P_{r1} - P_{r2}}{P_{r2} P_{r1}}}{\frac{P_{r1} - P_s}{P_s P_{r1}}} \\ &= \frac{P_s P_{r1} (P_{r1} - P_{r2})}{P_{r2} P_{r1} (P_{r1} - P_s)} = \frac{P_s P_{r1} - P_{r2} P_s}{P_{r2} P_{r1} - P_{r2} P_s} \end{aligned} \tag{3}$$

Where D_s is the distance between the front of the reference object and any other desired point where for example the subject might be found. If the assumption that the optical axis were parallel to the floor was violated, there would be an error in the computation of D_r. But there would be a proportional error in the calculation of D_s. Because these errors are proportional, the ratio between the two depths is still correct. Furthermore, because we use this correct ratio and our knowledge of the reference distance Dr, the computation for D_s is corrected. Because P_s is computed from P_{r1}, P_{r2} and the height differences H_r and H_s (see Figure 4), we can even cut short calculating P_s and D_r, and simplify to the following form:

[2]Zero roll is assumed. If any roll is determined then either the chosen lines should not be horizontal but instead follow the roll angle, or the picture should be turned upright first.

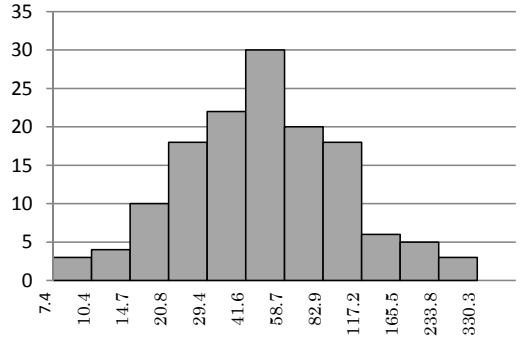


Figure 6: Observation counts per distance range in centimeters. Bin sizes increase logarithmically.

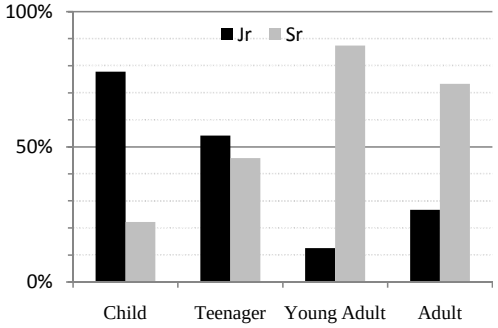


Figure 7: Observed interaction frequencies relative to robot type.

$$\frac{D_r}{D_s} = (\frac{H_r}{H_s} - 1)\frac{P_{r1}}{P_{r2}} + 1 \quad (4)$$

D_s will only provide a depth measurement though, we can combine this with a 'parallel plane' measurement to obtain a component perpendicular to the focal plane, and a component parallel to it. We can then use Pythagoras' theorem to determine the distance between any two points on the floor. Reference depth measurements (H_r) in the current experiment ranged roughly from 60 pixels to 450 pixels to capture a length of typically around 55 cm, giving almost centimeter precision or better.

5. RESULTS

140 Observations of 106 people were collected during a three day period. A Shapiro-Wilk test revealed that the data was not normally distributed. Inspection of the data suggested a logarithmic-normal distribution, which was confirmed by a second Shapiro-Wilk test on the logarithmically transformed data. All further tests were done on the transformed data. Out of the 140 observations one outlier was removed that was more than five standard deviations from the mean. The resulting transformed dataset had a mean of 3.87 and a standard deviation of 0.74. Subtracting or adding one standard deviation from the mean and transforming back to centimeters gives a 68.3% confidence distance interval of 23 to 100 cm with a mean of 48 cm (Table 3, Figure 6).

The natural logarithm of the chosen distance was analyzed using an analysis of variance with a $2 \times 2 \times 4 \times 4$, Robot type $\times$ Gender $\times$ Environment $\times$ Age group, unbalanced fractional factorial design. Since there were significant effects for Age group $\times$ Robot type [$F(3, 124) = 6.75, p < .0005$] and Age group $\times$ Gender [$F(3, 124) = 2.67, p = .05$], additional analyses were conducted per age group using a $2 \times 2 \times 4$ design. In no case did the environment reach significance. For children, robot type was significant [$F(1, 41) = 12.12, p = .001$]. As can be seen in Table3, the mean distance chosen by children was 26.8 cm for the small robot and 70.4 cm for the big robot. The Gender was a significant factor for teens [$F(1, 41) = 5.00, p = .03$] and marginally significant for adults [$F(1, 7) = 5.18, p = .057$]. The difference in chosen distance between male and female was remarkely large for Adults (93.5 versus 232.9 cm), but this difference was based on a few observations. For the Young Adults neither Robot type nor Gender were not significant factors in the chosen distance.

To test if subject height had any influence, the data set was split to age, but the age groups young adult and adult were pooled, since children and teens still grow and as such subject height is not an independent factor over all age groups. Since subject height would only be meaningful relative to robot height, separate tests were performed for Mobi Jr. and Sr. The main effect and the interaction with the environment were tested. For neither robot did subject height reach significance [$F(1, 5) = 2.68, p = .15$ for Mobi Jr.; $F(1, 38) = .002, p = .96$ for Mobi Sr.], nor did the interaction with environment [$F(1, 4) = 5.49, p = .08$ for Mobi Jr.; $F(1, 35) = 2.14, p = .11$ for Mobi Sr.].

Table 3 shows the mean distance in centimeters for the significant groups. Significant groups are bold and underlined. Since means and standard deviations were computed under logarithmic transformation, the distances these standard deviations represent are not equal in both directions. Therefore, the converted distances from one standard deviation below to one standard deviation above the mean centimeters are shown in brackets, providing a 68.3

Figure 7 shows the relative number of observations with each robot per age group. Since these observations are random in nature and not drawn from any distribution, no tests for significance can be performed. Children and teenagers are seen with Mobi Jr. respectively 3.5 and 1.2 times more often than with Mobi Sr. Young adults and adults are seen with Mobi Sr. respectively 7 and 2.8 times more often than with Mobi Jr.

6. DISCUSSION

All distances found in the present work except one suggest that the appropriate interaction distance for human-robot interaction lies within the personal zone of human interaction. The single divergent distance, which lies in the far phase or the social zone, is based on four observations, all of which show women watching the robot instead of talking to it. While this may be the preferred type of interaction for this group, the small number of observations is not sufficient to support this conclusion. Given the fact that this is the only incongruous result,

Table 3: Mean chosen distance in centimeters between subject and robot in different contexts. The 68.3% confidence interval (2 standard deviations) is shown in brackets. Significant results are underlined and bold.

	Child	Teenager	Young Adult	Adult	**All Ages**
Male	28.7 <15.5,53.0>	**39.9** **<23.7,67.0>**	57.1 <30.5,107.0>	**93.5** **<44.8,195.2>**	42.5 <20.6,88.0>
Female	33.3 <14.7,75.8>	**60.3** **<33.2,109.4>**	49.0 <33.7,71.4>	**232.9** **<176.6,307.1>**	53.6 <25.5,112.7>
Small Robot	**26.8** **<14.4,49.7>**	52.0 <29.5,95.0>	42.5 <42.5,42.5>	200.1 <117.1,342.4>	38.4 <16.9,87.2>
Big Robot	**70.4** **<36.7,135.1>**	55.1 <30.0,101.2>	53.8 <31.4,92.2>	91.4 <43.5,192.0>	58.4 <31.9,106.7>
All Groups	30.4 <15.1,61.1>	53.8 <29.5,98.0>	**47.6** **<20.3,111.6>**	126.7 <59.6,269.6>	47.9 <22.8,100.5>

there is reason to doubt the validity of this finding.

The personal distance found in the groups other than adult women is suitable for the type of interaction in this experiment among humans, and suggests acceptance of the robots as an agent that represents a social being. It should be noted however that in the case of Mobi Jr. it was apparent through conversations with it that people, especially children, did not always know it was controlled by a human being. In this case they could have accepted it as an autonomous agent that should be treated with similar social rules.

In the current work, the shape of the robot was only of influence on children. While this was in the line of expectation, since Mobi Jr. was specifically designed to work well with children, it was surprising to learn that other age groups made no distinction between the robots in choosing an interaction distance. There were however substantially more observations of children interacting with Mobi Jr. compared to Mobi Sr., and of young adults and adults interacting with Mobi Sr., indicating a preference of the respective age groups for those robots. While robots can be created with a myriad of possible appearances, it appears that the look of the robot is more important in appealing to a certain target audience than it is in influencing the preferred interaction distance. In this way, the appearance might be modeled with practical considerations in mind, such as the placement of sensors and visual or auditory outputs, or it might be made to resemble the target audience members, leading to a smaller cultural difference.

Instead of simply applying a set of learned norms to the robots, it is possible that people actually used similar criteria of sensory input that are mentioned in Section 1.2. In this context it could mean that the distance is chosen to facilitate communication. Practically this would mean standing close enough to hear the operator's voice through the speakers and to have the subject's own voice be picked up by the microphone which can be determined by the operator's communicated difficulty of hearing the subject. Mobi Sr. was shown at another exhibition where there was not enough light to see interaction partners through the webcam. A desk lamp was attached on top of its head, which influenced people's decisions on where to stand since people tended to step into the light. Perhaps audio manipulations such as loudness or stereo placement will show a similar influence on communication distance.

The distribution of chosen distances has been shown to be logarithmically normal. Although not necessarily logarithmic, a positively skewed distribution has been predicted by Sundstrom & Altman [7], and has been found in another human-robot interaction study by Walters et al. [8]. This means that in an approach starting from afar, comfort builds up slowly to an optimum and then drops off rapidly, possibly due to the undesirability of physical contact. Practically this means that if in doubt, it is better for a robot to stay back a bit too far rather than coming a bit too close, since overshooting the optimal distance will cause a much greater discomfort.

The average interaction distance of 47.9 cm is close to the verbal interaction distance of 62 cm reported by Koay et al. Note that our value for young male adults is even closer to the verbal interaction distance reported by Koay et al. [4]. However, the variance observed in this study is much larger than the variance previously reported: the differences in measured distances observed for human-robot proxemics studies is typically of the order of less than 20 cm. This is partly due to the effect of children interacting at close distance with the small robot, and female adults observing the robots from a far distance (note that these are independent observations, and is not explained by for example mothers watching their children interact). Yet, even for the Teenager and Young Adult groups, the variance was larger than previously reported, which is an indication of the variety of the audience attracted to this public event.

Surprisingly, the crowdedness of the environment is not significant in any of the groups. Having a surplus amount of space available to choose a position and communication distance was not expected to influence the choice, but given severe constraints people would still rather stand even closer to other people, than give up any space between themselves and the robot. There may be an alternative explanation however. Since the Mobi robots were a visitor attraction, people tended to crowd around them. This behavior led to spatial constraints for the people communicating with the robots at the front of the crowd. However, these subjects could have taken their preferred distance before the crowd limited them since people would gather behind or beside the subject not to disrupt his or her communication with the robot. Moreover, given the amount of space in the factory halls, there would typically be enough space around the crowd to provide everyone in it with at least personal distance. In-

vestigating communication distance between humans and robots in truly crowded environments would be difficult because of navigational problems. Perhaps human-robot distance preferences in such crowded conditions can be determined in an elevator setting, where there is no need for the robot to navigate through a crowd if it is the last one to exit and the first one to enter the elevator.

Pacchierotti et al. [6] describe a learning effect where comfortable distance becomes closer depending on whether a subject interacted with their robot in a previous trial. This could simply be caused by familiarity, but it might also be caused by a higher predictability of the robot's behavior which leads to a better estimate of whether or not the robot might be dangerous in any way. Apart from removing the need of keeping a cautionary distance, increased predictability and trust might also reduce the preferred interaction distance.

Additionally, a policy should be decided upon for dealing with learning effects. People may want to change their interaction with a certain robot or change their interaction distance as trust and familiarity is increased. To disregard initial cautious reactions on the human part would cause an unpleasant acquainting. On the other hand, to stay on the safe side and display solely more reserved manners might become a nuisance to frequent users. Ideally, robots should develop a social recognition system that determines whether or not any given person has been encountered before and what his or her attitude is towards the robot. However, such a system would normally not be available for all but the most advanced robots since the implementation of such a system is a far greater challenge than social distance maintenance.

7. CONCLUSIONS

It has been shown that age group is a significant factor in determining the preferred interaction distance, and furthermore that age group is of influence on what other factors play a role. Although the current work supports the notion that robot shape contributes mostly to appeal to a certain audience, it remains an open question if the shorter distances found in children's interaction with Mobi Jr. had practical grounds or were because of identification leading to a smaller cultural difference. Also, it remains unclear why there is a difference between the distance chosen by men and women in some age groups, whether or not this is related to cultural difference, and if the greater distance suggested for adult women is justified.

The influence of crowdedness and available space was not found to be significant in this work. Since the found preferred interaction distances were comparable to human personal and social space even when the environment provided enough capacity to keep public distance, there is no reason to doubt that any constraints that still provided the possibility to keep these distances were of any influence. In the case of intimate distance constraints however, the preferred distance would typically not be available without harming the preferred distance kept to other individuals. But in the current experiment, this space was available and the distance constraints were created only locally by crowding around the robot. Therefore, a further experiment is needed to establish the influence of severe spatial constraints in an environment that truly limits subjects to intimate distance.

8. REFERENCES

[1] J. C. Baxter. Interpersonal Spacing in Natural Settings. *Sociometry*, 33(4):444–456, December 1970.

[2] E. T. Hall. *The Hidden Dimension*. Anchor Books, New York, 1966.

[3] S. Heshka and Y. Nelson. Interpersonal Speaking Distance as a Function of Age, Sex, and Relationship. *Sociometry*, 35(4):491–498, December 1972.

[4] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn. Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction. In *16th IEEE International Conference on Robot and Human Interactive Communication*, pages 564–569, August 2007.

[5] S. Naidoo. Gender, Ethnicity, Intimacy and Proxemics. SARS 013 Class Paper, University of Pennsylvania, Fall 2000.

[6] E. Pacchierotti, H. I. Christensen, and P. Jensfelt. Evaluation of Passing Distance for Social Robots. In *15th IEEE International Workshop on Robot and Human Interactive Communication*, pages 315 – 320, September 2006.

[7] E. Sundstrom and I. Altman. Interpersonal Relationships and Personal Space: Research Review and Theoretival Model. *Human Ecology*, 4(1):47–67, January 1976.

[8] M. L. Walters, K. Dautenhahn, R. te Boekhorst, K. L. Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, and I. Werry. The Influence of Subjects' Personality Traits on Personal Spatial Zones in a Human-Robot Interaction Experiment. In *14th IEEE International Workshop on Robot and Human Interactive Communication*, pages 347 – 352, August 2005.

[9] F. Willes. Initial speaking distance as a function of the speakers' relationship. *Psychonomic Science*, 5:221–222, June 1966.