On Interaction Quality in Human-Robot Interaction

Suna Bensch[1], Aleksandar Jevtić[2] and Thomas Hellström[1]
[1]*Department of Computing Science, Umeå University, Umeå, Sweden*
[2]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain*
suna@cs.umu.se, ajevtic@iri.upc.edu, thomash@cs.umu.se

Keywords: Human-Robot Interaction, Evaluation, Performance

Abstract: In many complex robotics systems, interaction takes place in all directions between human, robot, and environment. Performance of such a system depends on this interaction, and a proper evaluation of a system must build on a proper modeling of interaction, a relevant set of performance metrics, and a methodology to combine metrics into a single performance value. In this paper, existing models of human-robot interaction are adapted to fit complex scenarios with one or several humans and robots. The interaction and the evaluation process is formalized, and a general method to fuse performance values over time and for several performance metrics is presented. The resulting value, denoted *interaction quality*, adds a dimension to ordinary performance metrics by being explicit about the interplay between performance metrics, and thereby provides a formal framework to understand, model, and address complex aspects of evaluation of human-robot interaction.

1 INTRODUCTION

HRI (Human-Robot Interaction) has several important similarities with HCI (Human-Computer Interaction). Models of interaction created for HCI have been applied and modified for HRI (Scholtz, 2003), and research in human factors for HCI are often applicable also in HRI (Young et al., 2011). As has been noted also by other authors (Scholtz, 2003; Fong et al., 2003; Young et al., 2011), HRI also differs from HCI in several important aspects. For the work in this paper, the following aspects are of particular interest: In HRI, several actors may take part, and interaction in all directions between humans, robots, and environment has to be considered. This is clearly emphasized in applications where robots collaborate with humans to accomplish a common mission, commonly known as Human-Robot Collaboration (HRC), such as assembling pieces of furniture (Dominey et al., 2009), or collaborative button-pressing (Breazeal and Hoffman, 2004). This highlights the need for updated models of interaction. Another difference between HCI and HRI is that the latter often involves several types of perception and action - at several levels, and with several modalities. The diversity of metrics and assessment methods adds further complexity, not least for social robots for which not only technical performance but also social aspects have to be considered. This highlights the need for a uniform and formal framework for description of interaction, and for how several metrics may be combined for evaluation of complex human-robot systems. In this paper we propose an updated model of interaction for HRI, taking into account the possibility of several humans and robots acting together. We also formalize the notion of interaction in the context of HRI, and use this as a basis for a suggested concept of *interaction quality*, that combines several performance metrics into one measure that indicates fitness of an interaction act for a given task. The results are expected to be valuable for both design and evaluation of HRI. In Section 2, earlier work on interaction models for HRI, and suggested metrics for evaluation of HRI are reviewed. In Section 3, an updated model of HRI is suggested. A formalism for interaction is presented in Section 4, followed by a computational derivation of the concept interaction quality in Section 5. Mechanisms for changes of interaction quality are discussed in Section 6, and Section 7 concludes the paper with a discussion and general conclusions.

2 EARLIER WORK

2.1 Modeling HRI

For a long time, robots were machines that humans gave explicit commands to in order to reach certain

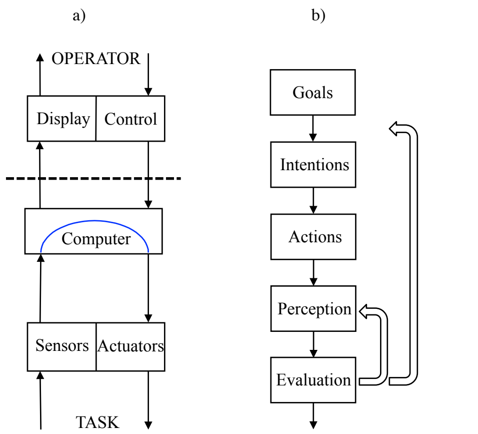

Figure 1: a) Model by (Sheridan, 1992), and b) Norman's modified model (Scholtz, 2003; Goodrich and Schultz, 2007) of interaction in a human supervisory control system.

goals. Such systems were appropriately modeled by simple feedback loops, such as the human supervisory control systems described in (Sheridan, 1992). Fig. 1a illustrates how such a robot, comprising a computer, sensors, and actuators, receives commands from a human operator through a communication channel (the dotted line) and a user interface comprising a display and a control unit. The robot may run at different levels of autonomy (illustrated by the arc), ranging from manual control through supervisory control, to fully automatic control. As autonomy increases, more responsibility is assigned to the computer, and less to the human, leading to decreased interaction between human and robot. Regardless of level of autonomy, human-environment interaction is modeled as mediated through the robot. While this is technically correct in the case of telerobotics (which was the main topic of the research in (Sheridan, 1992)), systems for telepresence (Minsky, 1980; Tsui et al., 2011) aim at creating the illusion of direct interaction, and the interaction models should therefore reflect this fact.

With the advent of more advanced robots, models from Human Computer Interaction (HCI) were adopted to the new field of Human-Robot Interaction (HRI). In (Scholtz, 2003), and also in (Barros and Lindeman, 2009), Don Norman's influential "gulf model" (Norman, 2002) was extended to describe various types of systems of humans and robots, in which the human may be supervisor, operator, bystander, or peer (Scholtz, 2003; Goodrich and Schultz, 2007). The example in Fig. 1b illustrates interaction between a robot and a human supervisor. Here interaction is considered at a higher level than the sensor and actuator data flow modeled in (Sheridan, 1992). Interaction is described in terms of goals, intentions, actions, perception, and evaluation. Most of the actions are automatically generated by the robot, and the main interaction between human and robot is the perception/evaluation loop, with additional interaction taking place at higher levels. In (Taha et al., 2011), Norman's model was adapted to HRI in another way by modeling interaction between human and robot with a POMDP (partially observable Markov decision process), and state variables intention, status, satisfaction and task. Such extensions are important, but the resulting models still focus on interaction between human and robot, or ignore (or abstract away) interaction with the environment. This is insufficient for a complete description of HRI, especially given insights from behavior-based robotics (Brooks, 1991), and other work (Pfeifer and Scheier, 2001; Ziemke, 2001; Brooks, 1990) emphasizing the importance of "embodiment", "situatedness" and robot-environment interaction.

2.2 Evaluating HRI

Several HCI evaluation techniques have been adopted and adapted to assess the efficiency and intuitiveness of HRI designs, such as Goal Directed Task Analysis (Adams, 2005), and Situation Awareness Global Assessment Technique (Endsley, 1988). The common approach with user testing in HCI has several specific drawbacks when applied to HRI (Drury et al., 2007). To overcome them, so called formal methods for modeling and evaluation have been developed in HCI, for instance the Goals, Operations, Methods, and Selection rules technique (GOMS) (Card et al., 1983; John and Kieras, 1996). GOMS has also been adapted to HRI (Drury et al., 2007) as a method to evaluate and predict efficiency of interfaces in HRI. While such methods are important tools to predict and evaluate user-related efficiency of interfaces for teleoperated and semiautonomous robots, they do not address the full range of performance metrics necessary for a complete evaluation of a complex human-robot systems. Several metrics for evaluation of systems comprising humans and robots working together have been suggested in the literature. An overview can be found in (Barros and Lindeman, 2009). Goodrich et al. (Goodrich and Olsen, 2003; Goodrich et al., 2003) describe principles and metrics focusing on effective and efficient interaction between teams of humans and robots. Special focus is on neglect tolerance, which quantifies for how long the robot can work unattended by the user. Steinfeld et al. (Steinfeld et al., 2006) argue that metrics to a large extent are task-dependent, and identify suitable metrics for tasks involving navigation, perception, management, manipulation, and social functionality. Common metrics for operator performance and robot are also suggested. For operator performance these metrics are situation awareness (see. e.g. (Scholtz et al., 2005)), workload, and accuracy of mental models of device operation. For robot performance, self-awareness, human awareness, and autonomy are listed as suitable metrics. For evaluation of the human experience of interacting with a social robot, the following metrics are suggested: interaction characteristics, persuasiveness, trust, engagement, and compliance. In (Kahn et al., 2007), the authors suggest a number of psychological benchmarks, or "... categories of interaction that capture conceptually fundamental aspects of human life ...": autonomy, imitation, intrinsic moral value, moral accountability, privacy, and reciprocity. These benchmarks can be used as additional metrics for evaluation of the experience of a human interacting with a social robot. In (Young et al., 2011) the authors propose three perspectives on social interaction in HRI: 1. Visceral factors of interaction, such as sadness, joy, frustration, and fear. 2. Social mechanisms including body language, and cultural norms. 3. Long-term social structures such as social relationships. These perspectives are suggested as basis for evaluation of human interaction experiences in HRI. Another analysis of metrics for social interaction is given in (Feil-Seifer et al., 2007; Feil-Seifer and Matarić, 2009), with emphasis on safety and scalability. To summarize, a large number of metrics have been suggested for evaluation of systems of robots working together with humans. In (Donmez et al., 2008), the lack of objective methods for selecting the most efficient metrics was recognized, and a list of evaluation criteria for metrics was presented. Still, there is a lack of a systematic and formal approach to how a large variety of metrics should be combined for evaluation of complex human-robot-environment systems.

3 A Model of HRI

To describe and analyze HRI involving in particular social robots, that work close to humans, we consider a model in which interaction is seen as an interplay between human(s), robot(s), and environment (see Fig. 2). Interaction takes place between all parts, and in all directions. In social robotics, interaction between human and robot goes well beyond pure technical sensor data and actuator commands, and comparisons with human-human interaction has been suggested (Krämer et al., 2012; Hiroi and Ito, 2012). A social robot may both perceive and express emotions such as joy, boredom, trust, and fear, and also other

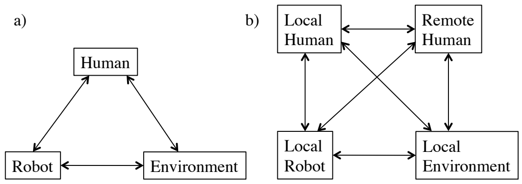

Figure 2: Proposed model of Human-Robot Interaction. a) Interaction between a human, robot, and environment located at the same place. b) A remote human added, for instance for telerobotics. In both cases, interaction with the environment is modeled as direct, although the exchange of information might be indirect via a communication link.

high-level properties such as tiredness and attention. Furthermore, interaction with such robots are not only about giving explicit commands that the robot then executes, and may also contain modes of interaction so far reserved for human-human interaction. For instance, a human may communicate verbally with a robot not by an imperative command such as "Bring me a glass of water", but by a declarative statement such as "I am thirsty". Hence, the arrows in Fig. 2 should be understood as interaction at a higher level than the exchange of sensor and actuator data referred to in the models of the kind illustrated in Fig. 1a. Fig. 2a shows a model with one human and one robot, for instance a domestic service robot. Several other combinations of single or multiple humans and robots are possible, where both humans and robots act as either individuals or in teams (Yanco and Drury, 2004). The model can, for such scenarios, be extended in a straightforward fashion. Fig. 2b shows a telepresence scenario, with one remote human controlling a local robot that interacts with a local human. Both humans and the robot are considered as agents, interacting with each other and the environment. Note that the remote human is considered to interact directly with the local human, robot, and environment, although the underlying physical information flow suggests that interaction between the local and remote side is mediated through the robot. This is an extension to earlier models that either do not include environment at all, such as in Fig. 1b and in (Scholtz, 2003; Olsen and Goodrich, 2003), or do not consider direct human-environment interaction, such as in Fig. 1a and in (Sheridan, 1992).

4 Formalizing Interaction

To describe and formalize interaction between agents we will introduce some definitions to break down the fuzzy concept of interaction. An interaction event e_t for a time interval t is defined as a tuple of perceived information and associated actions, for one or several agents. For instance, a human perceiving darkness and deciding to switch on the light may be regarded as an interaction event. Another example is a social robot that perceives that a human is sad and therefore suggests to call a relative. Yet another example is a human passing over on object to a robot. This example involves two agents perceiving and acting simultaneously during the same event.

An interaction act a is defined as an ordered sequence $[e_t, e_{t+1}, ...]$ of interaction events. For each interaction event, the decision process that leads to a chosen action may depend not only on the currently perceived data, but also on the perception history, the agent's state, prior knowledge, and of course the robot's general capabilities. Highly decisive is also the current goal or intention of the robot. For instance, a robot aiming at moving to a new location primarily chooses its action based on this navigational task.

A robot often makes a totally different assessment of a situation than does a human, and there are big differences also between humans. Even for one specific human, the way to act may depend on a huge number of time-varying factors such as mood, level of attention, level of tiredness, degree of happiness, anger, and even chance. Hence, several alternative interaction events are plausible for a given perceived information. For the continued analysis, we will use two examples of interaction, involving one robot and one or two humans interacting with each other. For a given time t, alternative interaction events are denoted $e_t^*, e_t^{**}, e_t^{***}$, etc.

Example 1:

A service robot R is programmed with the general goal of assisting an older male adult H with various domestic tasks in the apartment. H is sitting in his chair, feels a bit tired, and wants a cup of coffee. Table 1 lists some of several possible interaction events.

Table 1: Interaction events for Example 1.

Interaction event	Event description (percepts and actions)
e_1^*	H says "Robot, please help me make some coffee in the kitchen!", stands up and starts walking.
e_1^{**}	H stands up and starts walking.
e_2^*	R makes the assessment that H is tired (based on facial expression analysis) and slowly moves alongside H in order to provide physical support during the walk.
e_2^{**}	R does not make any assessment of H's level of tiredness (since the face of H is not visible) and moves after H towards the kitchen, in order not to be in the way.
e_2^{***}	R does nothing.
e_3^*	R asks H where he wants to go.

Four (of several possible) interaction acts based on these events are $a_1 = [e_1^*, e_2^*]$, $a_2 = [e_1^*, e_2^{**}]$, $a_3 = [e_1^*, e_2^*, e_3^*]$, $a_4 = [e_1^{**}, e_2^*, e_3^*]$, $a_5 = [e_1^*, e_2^{***}]$, $a_6 = [e_1^{**}, e_2^{***}]$. Several other interaction events and interaction acts exist with other combinations of R's perception of H. Note that a specific interaction event may occur for several different reasons. For instance, e_3^* can occur because R did not hear what H said (a_3), or because H did not say anything (a_4). Similarly, e_2^{***} may occur because R did not hear what H said (a_5), or H did not say anything (a_6).

Example 2:

A remote user H_r teleoperates a robot R within an environment containing a human H. Table 2 lists some of several possible interaction events. Two possible interaction acts based on these events are $a_1 = [e_1^*, e_2^*, e_3^*]$ and $a_2 = [e_1^{**}, e_2^*, e_3^{**}]$. Note that e_3^{**} is less likely to happen after e_1^* than after e_1^{**} since H_r in the former case has to focus on teleoperation and not on interacting with H (research indicates that teleoperation degrades performance for additional tasks (Chen, 2010; Glas et al., 2012)). The examples illustrate that entire interaction acts, and not individual interaction events, often have to be considered to be able to compare and evaluate interaction. For instance, the qualitative difference between e_1^* and e_1^{**} in Example 2 does not appear until two events later (e_3^* or e_3^{**}). Concerns related to how interaction events should be defined are discussed in Section 7.

Table 2: Interaction events for Example 2.

Interaction event	Event description (percepts and actions)
e_1^*	H_r teleoperates R towards H while asking H how he feels today.
e_1^{**}	H_r initiates autonomous navigation such that R moves by itself towards H. At the same time, H_r asks H how he feels today.
e_2^*	H tells H_r that everything is ok.
e_3^*	H_r asks H if he will watch TV.
e_3^{**}	H_r notices that H looks sad and asks him if he would like to join for shopping today.

5 Interaction Quality

To evaluate the fitness of an interaction act a for a specific task q, we introduce the term interaction quality denoted by the function $INQ(a,q)$. In this section we will investigate the nature of the function INQ.

To compute fitness, metrics such as those described in Section 2.2 are obviously useful. Most often, several perspectives have to be considered simultaneously. Robot performance such as speed and accuracy are often of primary concern, but also performance related to the experience of the involved humans are often considered important. For instance, situational awareness for an operator controlling a robot from distance, or the feeling of safety for a human having a robot as social companion may be relevant aspects to consider. In most cases, interaction with the environment plays an important role. Each perspective often needs several metrics for a thorough evaluation. Applying metric m on a given interaction act a gives us a performance value denoted $P(m)$. For a given task q with N required metrics we define the vector of all performance values:

$$\mathbf{P} = (P(1),...,P(N)). \tag{1}$$

Assessing fitness of an act requires considering all metrics at once. Hence, INQ can be written as a fusion function F that combines all performance values into a single measure:

$$INQ(a,q) = F(\mathbf{P}). \tag{2}$$

The task description q should include a specification of F, including information about the relative importance of all involved metrics. For some metrics, performance is most naturally expressed for each interaction event separately, and performance may vary from event to event. The influence of an individual interaction event on the performance of the entire interaction act sometimes depends on for how long the event is active. For instance, a short interaction event with high operator workload caused by low level of automation may be compensated by other events with higher level of automation and correspondingly lower operator workload. However, in some cases a very low performance during an interaction event disqualifies the entire interaction act, for instance in a navigation task in which the robot initially fails to perceive the instruction from the user on where to go. In the general case, a fusion mechanism f_m is needed to combine performance measured with metric m for individual interaction events into one measure for the entire interaction act (note that different fusion mechanisms may be necessary for different metrics, hence the subscript m):

$$P(m) = f_m(\mathbf{p}_m) \tag{3}$$

where

$$\mathbf{p}_m = (p_m(1),...,p_m(|a|)) \tag{4}$$

and $p_m(i)$ is the performance value for event i in interaction act a, computed with metric m. $|a|$ denotes the number of interaction events in interaction act a.

Note that $\mathbf{P}$, $P(i)$, $\mathbf{p}_m$, and $p_m(i)$ all depend on both q and a, but these dependencies have been left out in the notation for brevity. The computations and relations described above are graphically illustrated in Fig. 3. An interaction act a is divided into interaction events numbered $1,...,|a|$. For each metric m of the in total N metrics, performance measures $p_m(1),...,p_m(|a|)$ are computed at event level, and fused by the functions f_m to one measure $P(m)$ for the entire interaction act. The P values for all N metrics are finally fused by the function F to $INQ(a,q)$, the interaction quality for interaction act a when performing task q. Also the function F has to be defined in an task-dependent way. F and f_m may for instance be weighted averages, *min*, or *max* functions.

To summarize, performance values $p(i)$ for each metric i are first computed for all interaction events, and then fused over time to form $P(i)$. In this way, the temporal aspects of the interaction are considered. All $P(i)$ are then fused to form the overall interaction quality INQ. For some applications, an alternative formulation is more intuitive. In this formulation, fusion is first done over all performance values $p_1(i),...,p_N(i)$ for each interaction event i. These values are then fused with each other to form INQ. With reference to Fig. 3, this corresponds to having the functions f_m operating on columns instead of rows. With this formulation, the trade-off between for instance user friendliness and speed of execution at a specific instance in time may be more easily expressed. On the other hand, the trade-off between low and high performance, measured by the same metric for each event in the act, would be harder to express with this formulation. For practical implementation of the proposed framework, the choice of most suitable formulation depends on the specific task, and the preferred performance measures.

6 Discussion

Interaction quality depends on static and dynamic properties of the involved human(s), robot(s), and environment. As such it is relevant for both design and real-time operation of a robot. During iterative design of a robot, interaction quality should be a primary concern when evaluating selected hardware components and programmed basic behaviors. During operation of a robot, the effect on interaction quality

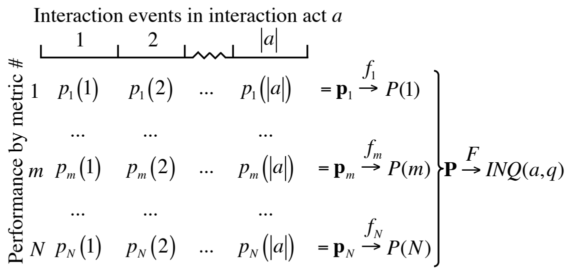

Figure 3: Derivation of interaction quality $INQ(a,q)$ for an interaction act a when performing task q. N metrics are used to compute performance related to both robot and human. Fusion of performance is done first at event, and then at metric level by the functions f_i and F respectively.

should be considered in all operations that influence the robot's actions and decision. This becomes particularly complex when considering the dynamic nature of an interaction system. The list below gives a few examples of changes in an interaction system that may influence interaction quality:
The human(s) may

- get tired, exhausted, impatient, or distracted.
- have a cup of coffee or take a nap.
- get distracted by other humans or environment.
- get scared of (or attracted by) the robot.

The robot(s) may

- decide to include a new sensor in the multi-modal perception system.
- change the level of autonomy, either under manual control or automatically.
- change relative priorities for tasks.

The environment is dynamic and

- typically changes in many respects as the human or robot moves to new locations.
- physical entities such as illumination, background noise, and floor conditions may change.
- external humans may start interacting with the human and/or robot.
- external humans may modify the environment, e.g. by moving pieces of furniture.

The influence of such internal and external changes is often intertwined, such that several performance measures are affected by the same change. Consider as an example a collaborative task with a robot emptying a dishwasher and a human placing the items in a cupboard. Increased background noise (change in the environment) will affect verbal interaction between the robot and the human. This leads to impatience and lowered human attention and also to misunderstandings in the robot's interpretations of the human's verbal commands. The result will be lowered performance measures for the human in terms of user satisfaction and for the robot in terms of task fulfillment. By properly designing the fusion functions f_i and F, such interdependencies can be identified and dealt with, both during design of the robot, and in real-time. Performance for varying settings can be simulated, and optimization procedures can be employed to maximize interaction quality.

Different performance metrics sometimes vary in different directions when a design parameter is changed. As an example, consider a teleoperation task in which an operator should navigate the robot in a fast and safe manner. The operator's workload typically increases as the speed of the robot increases, such that certain operator related performance measures decreases. At the same time, robot performance, measured by two metrics, vary in two directions as a function of speed; task achievement increases, and safety concern decreases. Another example is a human interacting verbally with a robot while walking on a slippery road. Just as in human-human interaction, the human may suddenly decide to focus on interacting efficiently with the environment in order not to fall. This causes performance related to communication between the human and the robot to decrease, and performance related to interaction between the human and the environment to increase. This examples illustrates how interaction quality often changes as a result of conscious decisions made and actions taken by the participating agents. Mechanisms that take changes of interaction quality into account may be included in the robot design, such that the robot changes behavior and character in order to maintain or increase overall interaction quality. One such example is sliding autonomy, in which the robot, autonomously or under human control, may increase its level of autonomy such that the remote user gets less involved in the physical control of the robot, and thereby can focus more on social interaction with the local user (Chen, 2010; Glas et al., 2012). In this case, the interaction quality changes due to several mechanisms. Performance related to human experience increases, while performance related to goal achievement for the robot decreases if the obstacle avoidance system slows down or even stops the robot in a crowded environment that a human operator would easily navigate through.

Granularity and appropriate chunking of events are essential for successful deployment of the proposed methodology. Interaction events can be defined at different levels of granularity. For the purpose of performance evaluation, events should be so short that further division into sub events would yield identical performance values for the sub events. For instance, event e_2^* in Example 1 above could in principle be divided into several shorter events since the robot's actions consist of a lot of low-level interaction with both H and the environment. However, interaction during these shorter events would be affected by design parameters and environment in a uniform way, and yield similar performance values as the undivided event. At the same time, an event should be so long that its performance values would be affected if the event would be further extended. For instance, e_2^* and e_3^* are affected by design parameters and environment in different ways, and merging these events into one would yield a different performance value.

7 Conclusions

The need to distinguish between performance at event and act level was discussed in (Olsen and Goodrich, 2003) and exemplified with a navigational task: "A robot might be getting closer to the target very rapidly and yet be wandering into a cul-de-sac from which it will need to back out". In such a case, performance at the event level will be very high, except for the events at which the robot has reached the dead-end. However, careful design of the fusion function f, should take the overall goal into account and result in a low overall performance at the act level. The proposed formalism highlights this kind of complex situations that may occur when evaluating performance of human-robot systems.

The introduced concept interaction quality is computed as a combination of several performance metrics. As such, it could be seen as nothing more than another way to compute performance for a human-robot system. However, we argue that the new concept adds a dimension by being explicit about the interplay between performance metrics. Fusion of performance is done over time by the functions f_i, and over metrics by the function F. In a complex system, possibly involving several robots and humans, a multitude of different performance metrics often have to be considered simultaneously. In many cases, they may be affected by modifying design parameters, and in many cases they vary due to changes in human and robot behavior, and in the environment. Often, trade-offs between conflicting performance measures have to be dealt with. Interaction quality and the proposed fusion mechanisms, working both over time and over different metrics, provide a formal framework to understand, model, and address these aspects of human-robot interaction. Apart from the general value of highlighting and formalizing these matters, the presented methodology may be implemented and used both for real-time optimization, and as a tool in design and evaluation of human-robot interaction. The practical value will be demonstrated and evaluated in real-world scenarios as part of future work.

ACKNOWLEDGEMENTS

This work was funded by the VINNOVA project 2011-01355, the EC project SOCRATES under MSCA-ITN-2016 grant agreement No 721619, and the SAMMIR project funded by CSIC, Spain, under the project reference number I-LINK1082.

REFERENCES

Adams, J. A. (2005). Human-robot interaction design: Understanding user needs and requirements. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 49, pages 447–451.

Barros, P. D. and Lindeman, R. (2009). A survey of user interfaces for robot teleoperation. Retrieved from: http://digitalcommons.wpi.edu/computerscience-pubs/21.

Breazeal, C. and Hoffman, G. (2004). Robots that work in collaboration with people. In *Proceedings of AAAI Fall Symposium on The Intersection of Cognitive Science and Robotics: From Interfaces to Intelligence*.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6:3–15.

Card, S., Moran, T., and Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Erlbaum, Hillsdale, New Jersey.

Chen, J. Y. C. (2010). Robotics operator performance in a multi-tasking environment. In Barnes, M., editor, *Human-Robot Interactions in Future Military Operations*. Ashgate.

Dominey, P. F., Mallet, A., and Yoshida, E. (2009). Real-time spoken-language programming for cooperative interaction with a humanoid apprentice. *International Journal of Humanoid Robotics*, 06(02):147–171, http://dx.doi.org/10.1142/S0219843609001711.

Donmez, B., Pina, P. E., and Cummings, M. L. (2008). Evaluation criteria for human-automation performance metrics. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, PerMIS '08, pages 77–82, http://dx.doi.org/10.1145/1774674.1774687, New York, NY, USA. ACM.

Drury, J. L., Scholtz, J., and Kieras, D. (2007). The potential for modeling human-robot interaction with GOMS. In

Sarkar, N., editor, *Human-Robot Interaction*, pages 21–38. I-Tech Education and Publishing.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *In Proceedings of the Human Factors Society 32nd Annual Meeting, Santa Monica, CA*.

Feil-Seifer, D. and Matarić, M. J. (2009). *Encyclopedia of Complexity and System Science*, chapter Human-Robot Interaction, pages 4643–4659. Springer.

Feil-Seifer, D., Skinner, K., and Matarić, M. J. (2007). Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3):432–439.

Fong, T., Thorpe, C., and Baur, C. (2003). Collaboration, dialogue, human-robot interaction. In Jarvis, R. A. and Zelinsky, A., editors, *Robotics Research: The Tenth International Symposium*, pages 255–266, http://dx.doi.org/10.1007/3–540–36460–9_17, Berlin, Heidelberg. Springer Berlin Heidelberg.

Glas, D., Kanda, T., Ishiguro, H., and Hagita, N. (2012). Teleoperation of multiple social robots. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 42(3):530–544, http://dx.doi.org/10.1109/TSMCA.2011.2164243.

Goodrich, M. A., Crandall, J. W., and Stimpson, J. L. (2003). Neglect tolerant teaming: Issues and dilemmas. In *Proceedings of the 2003 AAAI Spring Symposium on Human Interaction with Autonomous Systems in Complex Environments*.

Goodrich, M. A. and Olsen, D. R. (2003). Seven principles of efficient human robot interaction. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 4, pages 3942–3948 vol.4, http://dx.doi.org/10.1109/ICSMC.2003.1244504.

Goodrich, M. A. and Schultz, A. C. (2007). Human-robot interaction: A survey. *Found. Trends Hum.-Comput. Interact.*, 1(3):203–275, http://dx.doi.org/10.1561/1100000005.

Hiroi, Y. and Ito, A. (2012). *Informatics in Control, Automation and Robotics: Volume 2*, chapter Toward Human-Robot Interaction Design through Human-Human Interaction Experiment, pages 127–130, http://dx.doi.org/10.1007/978–3–642–25992–0_18. Springer Berlin Heidelberg, Berlin, Heidelberg.

John, B. E. and Kieras, D. E. (1996). Using GOMS for user interface design and evaluation. *ACM Transactions on Human-Computer Interaction*, 3(4).

Kahn, P. H., Ishiguro, H., Friedman, B., Freier, N. G., Severson, R. L., and Miller, J. (2007). What is a human? toward psychological benchmarks in the field of human-robot interaction. interaction studies: Social behaviour and communication in. *Biological and Artificial Systems*, pages 1–5.

Krämer, N. C., Pütten, A., and Eimler, S. (2012). *Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction*, pages 215–240, http://dx.doi.org/10.1007/978–3–642–25691–2_9. Springer Berlin Heidelberg, Berlin, Heidelberg.

Minsky, M. (1980). Telepresence. *Omni*, 2(9):45–52.

Norman, D. A. (2002). *The Design of Everyday Things*. Basic Books, Inc., New York, NY, USA.

Olsen, D. R. and Goodrich, M. A. (2003). Metrics for evaluating human-robot interaction. In *Proceeedings of NIST Performance Metrics for Intelligent Systems Workshop*.

Pfeifer, R. and Scheier, C. (2001). *Understanding Intelligence*. MIT Press.

Scholtz, J. (2003). Theory and evaluation of human robot interactions. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 10–, http://dx.doi.org/10.1109/HICSS.2003.1174284.

Scholtz, J., Antonishek, B., and Young, J. (2005). Implementation of a situation awareness assessment tool for evaluation of human-robot interfaces. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 35(4):450–459, http://dx.doi.org/10.1109/TSMCA.2005.850589.

Sheridan, T. B. (1992). *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, MA, USA.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M. A. (2006). Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI06, pages 33–40, http://dx.doi.org/10.1145/1121241.1121249, New York, NY, USA. ACM.

Taha, T., Miro, J., and Dissanayake, G. (2011). A POMDP framework for modelling human interaction with assistive robots. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 544–549, http://dx.doi.org/10.1109/ICRA.2011.5980323.

Tsui, K. M., Desai, M., Yanco, H. A., and Uhlik, C. (2011). Exploring use cases for telepresence robots. In *Proceedings of the 6th International Conference on Human-robot Interaction*, HRI '11, pages 11–18, http://dx.doi.org/10.1145/1957656.1957664, New York, NY, USA. ACM.

Yanco, H. and Drury, J. (2004). Classifying human-robot interaction: an updated taxonomy. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 3, pages 2841–2846 , http://dx.doi.org/10.1109/ICSMC.2004.1400763.

Young, J. E., Sung, J., Voida, A., Sharlin, E., Igarashi, T., Christensen, H., and Grinter, R. (2011). Evaluating human-robot interaction: Focusing on the holistic interaction experience. *International Journal on Social Robotics*, 3(1):53–67.

Ziemke, T. (2001). Are robots embodied? In Balkenius, Zlatev, Dautenhahn, Kozima, and Breazeal, editors, *Proceedings of the First International Workshop on Epigenetic Robotics - Modeling Cognitive Development in Robotic Systems*, volume 85, pages 75–83, Lund, Sweden. Lund University Cognitive Studies.