Applying Heuristic Evaluation to Human-Robot Interaction Systems

Edward Clarkson and Ronald C. Arkin

College of Computing and GVU Center, Georgia Institute of Technology
801 Atlantic Drive, Atlanta, GA 30332-0280
{edcclark, arkin}@cc.gatech.edu

Abstract

Though attention to evaluating human-robot interfaces has increased in recent years, there are relatively few reports of using evaluation tools during the development of human-robot interaction (HRI) systems to improve their designs. Heuristic evaluation is a technique suitable for such applications that has become popular in the human-computer interaction (HCI) community. However, it requires usability heuristics applicable to the system environment. This work contributes a set of heuristics appropriate for use with HRI systems, derived from a variety of sources both in and out of the HRI field. Evaluators have successfully used the heuristics on an HRI system, demonstrating their utility against standard measures of heuristic effectiveness.

Introduction

The attention paid to human-robot interaction (HRI) issues has grown dramatically as robotic systems have become more capable and as human contact with those systems has become more commonplace. Along with the development of robotic interfaces, there has been an increase in the evaluation of these systems. HRI researchers can employ a variety of evaluation styles in their work; they can evaluate their systems *summatively* (i.e., after-the-fact) or *formatively* (i.e., during system development). However, there have been relatively few accounts of formative applications or uses of *discount* (low-cost) techniques. Discount methods used in formative evaluations can be powerful tools. Not only do they take small amounts of time and resources, but they can catch both major and minor problems early in the development cycle.

One discount evaluation technique is heuristic evaluation (HE) (Nielsen and Molich 1990; Nielsen 1994), a method that has become popular in the HCI community. HE consists of a small group of evaluators who examine an interface using a set of heuristics as a guide for their inspection. Its low cost makes it well suited to formative evaluations, and studies have shown its effectiveness relative to other discount approaches (Jeffries et al. 1991). But the application of HE to a problem requires heuristics that are appropriate for the problem domain.

This work presents our efforts to synthesize such a set of HRI-specific heuristics. This allows for application of HE to HRI systems and also encourages the use of formative evaluations in HRI system design. Our development procedure is based on accepted methodology from previous adaptations of heuristic evaluation to new problem domains (Baker, Greenberg and Gutwin 2002, Mankoff et al. 2003), and takes inspiration for the heuristics themselves from a variety of existing works in HRI (Goodrich and Olsen 2003, Scholtz 2002, Sheridan 1997, Yanco, Drury and Scholtz 2004) and related fields (Baker, Greenberg and Gutwin 2002, Mankoff et al. 2003, Nielsen 1994). We present an application of our derived heuristics to the evaluation of a sample HRI system, which shows that 3-5 evaluators using the set find 40-60% of known usability problems (the standard test for heuristic effectiveness). We also find no statistically significant difference in the quantities or the severity of the problems found by HCI and robotics evaluators. The result is a validated set of HRI heuristics, suitable for use by roboticists with little or no previous evaluation experience.

Related Work

Evaluation on HRI systems has not received its due attention until recently. A recent DARPA/NSF report, for example, cites the need for methods and metrics that can be used to measure the development of human-robot teams (Burke et al. 2004). A common approach among evaluations reported in the literature has been field-based case studies of various systems (Nourbakhsh et al. 1999; Schipani 2003; Yanco, Durry and Scholtz 2004). The number of controlled lab studies of HRI interfaces has also increased over recent years, with most focused on comparing interface alternatives for teleoperation (Johnson, Adams and Kawamura 2003; Olivares et al. 2003) or plan specification software (Endo, MacKenzie and Arkin 2004).

Formative and Summative Evaluation

A common thread—with a few exceptions (Nourbakhsh et al. 1999)—among HRI evaluation literature is the focus on summative studies and techniques. Summative evaluations judge the outcome of a design implementation while formative evaluations assess preliminary design products. Note that this distinction concerns when a technique is applied and not the method *per se*. We have already noted that research is needed into techniques that can be used to judge the progress of HRI systems; we make two observations about the state-of-the-art in HRI:

- There are relatively few validated tools and techniques for evaluating HRI systems.
- There have been few reports of performing formative evaluations (or other principles of user-centered design) in the HRI literature (Adams 2002).

Cleary, the former contributes to the latter. Hence, there is a need for evaluation methods that are suited to formative studies and have been demonstrated on HRI applications.

Discount and Heuristic Evaluation

Discount evaluation techniques are methods designed explicitly to be low cost (in terms of manpower and time). Because of these properties, discount evaluations such as heuristic evaluation are often applied formatively. HE is applicable to a wide range of prototypes, from design specifications to functioning systems, and has been empirically validated (Nielsen and Molich 1990, Nielsen 1994, Jeffries et al. 1991). It requires few (three to five) evaluators who need not be domain experts (though it is more effective with such training).

The principle behind HE is that individual usability inspectors of a system do a relatively poor job, finding a fairly small percentage of the total number of known usability problems. However, Nielsen has shown evaluators have a wide variance in the problems they find, which means the results of a small group of evaluators can be aggregated with little duplication to uncover a large number of bugs. Briefly, the HE process in particular consists of the following steps:

- Pre-evaluation preparation:
    1. Create problem report templates for evaluators.
    2. Customize heuristics to the specific interface being evaluated: certain heuristics may be irrelevant to the evaluation goals, and heuristic descriptions given to the evaluators can include references and examples taken from the system in question.
- Assemble a small group of evaluators (Nielsen recommends three to five) to perform the HE.
- Each evaluator **independently** assesses the system in question and judges its compliance with a set of usability guidelines (the heuristics).
- Either the evaluators or the experimenter aggregate the results from each evaluator and assign severity ratings to the various usability issues generated.

HE has been shown to find 40 – 60% of usability problems with just three to five evaluators. Of course, the results of an HE are not comprehensive, highly subjective and not repeatable. However, it should be emphasized that this is not a goal for HE. Its purpose is to provide a evaluation framework that is easy to teach, learn and perform while also uncovering significant numbers of usability problems early in the design process.

Heuristic Evaluation for HRI

The problem with applying heuristic evaluation to HRI, however, is the validity of using existing heuristics for HRI. Nielsen's standard heuristics have been directly applied to HRI systems (Drury et al. 2003), but we are not aware of any heuristics that have been compiled specifically for the domain of HRI. Hence, an obvious question is whether it is possible to form a new set of heuristics that are pertinent to HRI systems?

If we rely on the HCI literature, the answer is "yes." Alternative heuristic sets have already been developed for domains outside on-the-desktop software. Confronted with similar problems—Nielsen's heuristics do not focus on computer-supported cooperative work (CSCW) issues—researchers adapted heuristics for use with groupware applications (Baker, Greenberg and Gutwin 2002). Similarly, Mankoff and her collaborators produced a heuristic list for ambient displays[1] (Mankoff et al. 2003). In each case, researchers developed new heuristic lists and validated them using similar methodologies.

The essence of the process in both of these works is to generate an initial list of heuristics (via brainstorming and related work); employ them in an evaluation; analyze the results; iterate. The consensus benchmark (from Nielsen) is that an average group of three to five inspectors should find between 40% and 60% of known usability problems.

The problem domains for each of these adaptations are noteworthy. CSCW applications are concerned with facilitating teamwork, organizing group behavior and knowledge, and providing support for concurrent interaction—all items relevant to HRI. Likewise, ambient devices convey information with low attentional requirements, and do so through a variety of form factors. HRI systems face similar issues in trying to maintain operator awareness of sensor data or human response to a robot's physical appearance.

HRI Heuristic Development

Following the methodology similar to Baker et al. and Mankoff et al., (which were based in turn on Nielsen's method for creating his initial list), our process for heuristic development consists of three broad steps: create an initial list of HRI heuristics via brainstorming and synthesizing existing lists of potentially applicable heuristics; modify the list based on pilot studies, consultation with other domain experts, and other informal techniques; and validate the modified list against an existing HRI system.

There are a number of bases from which to develop potential HRI heuristics: Nielsen's canonical list (Nielsen 1994); HRI guidelines suggested by Scholtz (Scholtz 2002); and elements of the ambient and CSCW heuristics (Baker, Greenberg and Gutwin 2002; Mankoff et al. 2003). Sheridan's challenges for the human-robot communication

[1] *Ambient displays* provide information at the fringe of users' attention, and are usually visually appealing.

(Sheridan 1997) can also be considered issues to be satisfied by an HRI system.

These lists and the overall body of work in HRI provide the basis for heuristics applicable to both multi-operator and multi-agent systems; however, for this work we have limited our focus to single operator, single agent settings. This narrows the problem focus, so lessons learned during this initial work can be applied to further development.

Our initial list is based on the distinctive characteristics of HRI, and should ideally be pertinent to systems ranging from windowed software applications to less traditional interfaces (e.g., the iRobot Roomba vacuum cleaner). Our list should also apply equally well to purely teleoperated machines and monitored autonomous systems.

Norman emphasizes a device's ability to communicate its state to the user as an important characteristic (Norman 1990). Applied to a robot, an interface then should make evident various aspects of the robots status—what is its pose? What is its current task or goal? What does it know about its environment, and what does it not know? Parallel to the issue of *what* information should be communicated is *how* it is communicated. The complexity of sensor data is such that careful attention is due to what the user needs out of that data, and designing an interface to convey that data in its most useful format.

Many of these questions have been considered as a part of the heuristic sets mentioned previously, and we leverage that experience by taking elements in whole and part from those lists to form our own attempt at an HRI heuristic set. Since the heuristics are intended for HRI systems, they focus only on the characteristics distinct to HRI. The inspirational source or sources before adaptation accompany each heuristic in Table 1. Space constraints prevent us listing the initial heuristic descriptions; we refer the reader to an expanded version of this paper for those descriptions and other commentary (Clarkson and Arkin 2006). Heuristics 1, 2, and 3 all deal with the handling of information in an HRI interface. Heuristics 4, 5 and 6 all deal with the form communication takes between the user and system and vice versa. Heuristic 5 is indicative of an interface's ability to immerse the user in the system, making operation easier and more intuitive. Heuristic 7 reflects the longevity and adaptability often required of HRI platforms. Finally, Heuristic 8 signifies the potential importance of emotional responses to robotic systems.

HRI Heuristic Validation

Our validation plan is similar to that described in both Baker and Mankoff:

- Create an initial list of HRI heuristics via brainstorming and synthesizing existing lists of potentially applicable heuristics (see above).
- Use the heuristics in an evaluation of an HRI system.
- Hypothesize that a small number of evaluators using the heuristics will uncover a large percentage of known usability problems.
- Modify the initial heuristic list based on the results.

1. **Sufficient information design** *(Scholtz, Nielsen)*
2. **Visibility of system status** *(Nielsen)*
3. **Appropriate information presentation** *(Scholtz)*
4. **Match between system and real world** *(Nielsen, Scholtz)*
5. **Synthesis of system and interface** *(None)*
6. **Help users recognize, diagnose, and recover from errors** *(Nielsen, Scholtz)*
7. **Flexibility of interaction architecture** *(Scholtz)*
8. **Aesthetic and minimalist design** *(Nielsen, Mankoff)*

Table 1 – Initial HRI heuristic titles.

It is necessary to have a relatively large group of evaluators for the purposes of assessing the heuristics. Though HE generally requires only a few (3-5) evaluators, a larger group lets us test whether an arbitrary subset of the overall group can indeed uncover a significant percentage of usability problems.

Experimental Method. We lacked an active project using an HRI system appropriate for such an evaluation, so we created an *ad hoc* system and problem for this work. We chose the RoboCup Rescue, an annual worldwide robotics competition, as our problem environment. The contest is held in an indoor arena designed to mimic a portion of an post-disaster urban area. We chose a robot based on the Segway RMP platform as the HRI system to be evaluated. The system as presented to users was teleoperated using the Mobile Robot Lab's MissionLab software package and a standard PC analog joystick controller. It contained two major sensory systems: a pair of SICK laser rangefinders (mounted parallel to the floor) and a forward-mounted optical camera.

We have presented only an outline of the system in question; the contribution of this work is not the system but the results of that evaluation only as it informs the development of our heuristics. Indeed, our HRI system and problem environment are not particularly well-suited for each other by design. The purpose of a HE is to uncover interaction problems (the more severe the better), and a problem/system mismatch ensures their presence for evaluators. We report the specificities of the problems indicated by our evaluation only insofar as they inform the development and validation of our heuristics.

We recruited ten HCI and robotics graduate students to serve as our evaluator team. Two did not complete the entirety of the evaluation and are ignored henceforth. The eight remaining (five female) had a mean age of 28 years. Three evaluators had a specialization in robotics and the other five specialized in HCI.

We prepared and distributed a packet of written information summarizing the experiment and its goals ahead of an introductory meeting. The packet included an introduction and summary of both the HRI system and the problem environment, and copies of problem report templates. The templates provided fields for a problem title, detailed description, and an indication of which heuristic the problem violated. We also discussed the information contained in the packet in a meeting with the evaluators. This included an introduction to the heuristic evaluation procedure, a presentation on the robot's sensors and capabilities, the RoboCup rescue competition and rules, and a live demonstration of the operation of the system. Evaluators were encouraged to return within a week as many problem reports as they deemed appropriate. We also instructed them to prepare their problem reports independently. We did not suggest specific time-on-task guidelines for completing the problem reports.

Results

The evaluators as a group returned 59 problem reports. Individual counts ranged from 5 to 10. We synthesized the results by combining duplicate problem reports. Such duplicates were sometimes obvious ("Only one camera; doesn't move; doesn't cover 360 deg." and "Camera direction/control [is] fixed position, can't move it.") Other duplications were more subtle: "Map doesn't show orienting features, can't mark locations of interest. No ability to save history of movements" and "Need indication of how many victims found and where, hazards and locations, running point total" reflect different aspects of the same problem (the system does not effectively provide historical data about significant environmental features).

After synthesizing the results, we identified 21 unique problems and assigned severity ratings to each of them using a standard rating system of 0-4, with 4 being the most severe and 0 being a non-problem. Evaluators found 11 severe problems (ratings or 3 or 4) and 10 minor problems (ratings of 1-2). Average severity across all 21 problems was 2.52. There were no non-problems reported. The average single evaluator found 29% of the known problems, a figure comparable to other reports (Baker, Greenberg and Gutwin 2002; Nielsen and Molich 1990). Figure 1 shows a representation of how problem identification is distributed across the different evaluators. Evaluators are represented by rows and ordered from least to most successful (measured by the number of unique problems reported). Each column signifies a unique problem, and they are ordered according to severity. The chart shows that there is substantial spread among the different evaluators and that even relatively unsuccessful evaluators are able to identify severe problems.

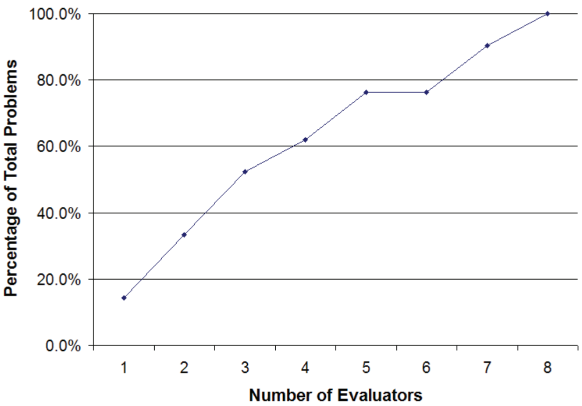

Figure 1 – Percentage of known problems found with increasing number of evaluators (ordering is random).

Similarly, Figure 1 charts the increasing percentage of known problems found with additional evaluators (with additions to the total from the least to most successful evaluator). Most notably, it shows that the heuristics passed the canonical HE test: 3-5 evaluators identify at least 40-60% of the known problems. An inspection of other randomly-ordered graphs showed similar results.

Many robotics projects have limited access to HCI specialists, or at least have easier access to roboticists. As a result, we were also interested if there are any differences between evaluators with a background in robotics (*group R*) and HCI (*group H*). In our evaluation, group H found 7.6 unique problems against a mean of 5.3 for group R, a difference which is marginally significant (2-tailed t; p = 0.06). The average severity of the problems found by each group was almost identical at 2.53 for group H and 2.51 for group R. Notable is the fact that all of group H had participated in and 80% had themselves conducted an HE prior to our study; only one member of group R had participated in or conducted an HE. As such, familiarity with the HE process may be a contributing factor to this result. However, even with apparently less effective evaluators, the three roboticists identified 43% of the total

	<< Less Severe							Problems ordered by severity							More Severe >>						
Evaluators ordered by success << less more >>						■		■		■	■	■		■					■	■	■
	■		■	■	■		■	■							■		■		■		
							■		■	■						■	■			■	■
										■				■				■	■	■	■
		▨	▨		▨											▨				▨	
					▨		▨				▨		▨			▨					
								■	■	■							■				■
								▨			▨									▨	

Table 2 – A chart of the problems found by each evaluator. A filled square indicates the problem corresponding to that row was identified by the evaluator corresponding to that column. Black rows are HCI specialists; gray rows are roboticists.

problems, still within the standard for an acceptable HE process. This indicates that teams of roboticists can perform effective HEs with little or no prior experience.

Discussion

A number of issues with our heuristics arose explicitly via evaluator comments or implicitly through their problem responses. One of the most severe problems with our example HRI system is that its sensor capabilities simply are not adequate to perform the tasks expected in the competition. However, none of our heuristics plainly mention checking system capabilities against expected tasks (though heuristic 7 comes close). Likewise, many of the other most severe problems with our HRI system relate to the difficulty in maintaining an accurate mental model of the robot and its surroundings (often termed *situational awareness*). Heuristic 2 touches on this idea, but does not use the *situational awareness* term explicitly.

We also found our use of an *ad hoc* system for evaluation purposes to be limiting in some ways. Since we did not employ the system for its purported use (i.e., compete in RoboCup Rescue), we cannot have a true appreciation for the full scope of the problems. Similarly, because there were many obvious mismatches between the task and our HRI system, it is difficult to gauge whether the existing problems could have been qualitatively different from ones in a more realistic scenario.

We should also include some general discussion of when the HE approach is *not* suitable. We have noted HE results have no guarantee of repeatability or comprehensiveness, and are not amenable to statistical analyses. Thus, HE is not a good choice for measuring system performance or comparing systems: its utility is practically limited to guiding the formative stages of system development via early, iterative applications.

Goodrich and Olsen have also proposed seven principles for effective HRI systems (Goodrich and Olsen 2003). They are: implicitly switch interfaces and autonomy modes; let the robot use natural human cues; manipulate the world instead of the robot; manipulate the relationship between the robot and world; let people manipulate presented information; externalize memory; and help people manage attention. Many of these principles are covered explicitly or implicitly in our initial heuristic set, though they were not used in their original development. For example, "use natural cues" is similar to heuristic 4. To "directly manipulate the world" requires an interface which acts as an extension of the HRI system (heuristic 5).

1. **Sufficient information design.**

   The interface should be designed to convey "just enough" information: enough so that the human can determine if intervention is needed, and not so much that it causes overload.

2. **Visibility of system status**

   The system should always keep users informed about what is going on, through appropriate feedback within reasonable time. The system should convey its world model to the user so that the user has a full understanding the system's situational awareness.

3. **Appropriate information presentation**

   The interface should present sensor information that is easily understood and in a useful form. The system should use the principle of recognition over recall, externalizing memory and improving users' situational awareness via attention management.

4. **Use natural cues**

   The language of the interaction between the user and the system should be in terms of words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

5. **Synthesis of system and interface**

   The interface and system should blend together so the interface is an extension of the user, the system and by proxy, the world. The interface should facilitate efficient communication between system and user, switching modes automatically when necessary.

6. **Help users recognize, diagnose, and recover from errors**

   System malfunctions should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution. The system should present enough information about the task environment so that the user can determine if some aspect of the world has contributed to the problem.

7. **Flexibility of interaction architecture**

   If the system will be used over a lengthy period of time, the interface should support the evolution of system capabilities, such as sensor and actuator capacity, behavior changes and physical alteration. Sensor and actuator capabilities should be adequate for the system's expected tasks and environment.

8. **Aesthetic and minimalist design**

   The physical embodiment of the system should be pleasing in its intended setting. The system should not contain information that is irrelevant or rarely needed.

Table 3 – Revised HRI heuristics.

Updated HRI Heuristics

Though our heuristics performed well in our tests, our findings led us to revise our heuristics, clarifying them by rewording or adding various passages. The final results are presented in Table 3. We have added an overt mention of situational awareness to heuristics 2 and 3; changed heuristics 3 and 5 to reflect better several of Goodrich and Olsen's principles; re-titled heuristic 4 with their "use natural cues" phrase, which is clearer and more succinct than the original heading; and changed heuristic 7 to ensure a check for appropriate hardware capabilities.

Future Work and Conclusions

Future work in this area is promising. Certainly, iterative use of the heuristics is key to their improvement. Their indirect promotion of formative evaluation can improve the efficiency and efficacy of HRI development efforts. Their continued use may also inform the heuristic development for multi-robot or -human systems.

The utility of formative evaluations is strong motivation for the use of such methods in HRI. Heuristic evaluation, a usability inspection method from HCI, is ideal for formative applications. Previous work has validated the concept of adapting HE to new problem domains. We have proposed an initial set of heuristics intended for single operator, single agent human-robot interaction systems, validated them against an example HRI system and amended the set based on our experience. Our tests also indicate no significant differences between robotics and HCI evaluators, indicating teams of roboticists can independently perform successful HEs.

Acknowledgements

Thanks are due to our volunteer evaluators for their generous contributions of time and effort, and to Yoichiro Endo for his review of an early draft of this work.

References

Adams, J. 2002. Critical Considerations for Human-Robot Interface Development. In *Proceedings of the 2002 AAAI Fall Symposium on Human-Robot Interaction*, 1-8.

Baker, K., Greenberg, S. and Gutwin, C. 2002. Empirical development of a heuristic evaluation methodology for shared workspace groupware. In *Proceedings of CSCW '02*, 96–105.

Burke, J., Murphy, R.R., Rogers, E., Scholtz, J., and Lumelsky, V. 2004. Final Report for the DARPA/NSF Interdisciplinary Study on Human-Robot Interaction. *IEEE Systems, Man and Cybernetics* C 34 (2): 103-112.

Clarkson, E. and Arkin, R. 2006. Applying Heuristic Evaluation to Human-Robot Interaction Systems. GVU Tech Report No. GIT-GVU-06-08, ftp://ftp.cc.gatech.edu/pub/gvu/tr/2006/06-08.pdf

Drury, J., Riek, L., Christiansen, A., Eyler-Waler, Z., Maggi, A., and Smith, D. 2003. Command and Control of Robot Teams. In *Proceedings of AUVSI '03*.

Endo, Y., MacKenzie, D.C., and Arkin, R. 2004. Usability Evaluation of High-Level User Assistance for Robot Mission Specification. *IEEE Transactions on Systems, Man, and Cybernetics* C, 34 (2): 168-180.

Goodrich, M. and Olsen, D. 2003. Seven Principles of Efficient Human Robot Interaction. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetic '03*, 3943-3948.

Jeffries, R., Miller, J., Wharton, C. and Uyeda, K. User Interface Evaluation in the Real World: a Comparison of Four Techniques. 1991. In *Proceedings of CHI '91*, 119-124.

Johnson, C., Adams, J. and Kawamura, K. Evaluation of an Enhanced Human-Robot Interface. 2003. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics '03*, 900-905.

Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Ames, M., Lederer, S. 2003. Heuristic evaluation of ambient displays. In *Proceedings of CHI '03*, 169-176.

Nielsen, J. and Molich, R. 1990. Heuristic Evaluation of User Interfaces. In *Proceedings of CHI '90, 249-256.*

Nielsen, J. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of CHI '94*, 152-158.

Norman, D. 1990. *The Design of Everyday Things*. Doubleday, New York, 1990.

Nourbakhsh, I., Bobenage, J., Grange, S., Lutz, R., Meyer, R. and Soto, A. 1999. An affective mobile robot educator with a full-time job, *Artificial Intelligence* 114 (1-2): 95–124.

Olivares, R., C. Zhou, J. Adams, and B. Bodenheimer. 2003. Interface Evaluation for Mobile Robot Teleoperation. In *Proceedings of the ACM Southeast Conference '03*, 112-118.

Schipani, S. 2003. An Evaluation of Operator Workload, During Partially-Autonomous Vehicle Operation. In *Proceedings of PERMIS '03.*

Scholtz, J. 2002. Evaluation methods for human-system performance of intelligent systems. In *Proceedings of PERMIS '02*.

Sheridan, T. 1997. Eight ultimate challenges of human-robot communication. In *Proc. of RO-MAN '97*, 9–14.

Yanco, H., Drury, J. and Scholtz, J. Beyond Usability Evaluation: Analysis of Human-Robot Interaction at a Major Robotics Competition. *Journal of Human-Computer Interaction*, 19 (1-2): 117-149.