Evaluating Human-Robot Interaction in a Search-and-Rescue Context[*]

Jill Drury, Laurel D. Riek, Alan D. Christiansen,
Zachary T. Eyler-Walker, Andrea J. Maggi, and David B. Smith
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102 USA
{jldrury, laurel, adc, zach, amaggi, daves}@mitre.org

ABSTRACT

We have designed a novel interface to enable command and control of a team of mobile robots by a single operator. To understand the usability of MITRE's human-robot interface, we have embarked on a multi-step evaluation program. This paper describes the first step in the program, which is to apply techniques from the human-computer interaction field to evaluate how well the interface supports a single user controlling a single robot in an urban search-and-rescue environment. In particular, this paper describes our attempts to evaluate the interface in terms of external measures that benchmark the ability of a human-robot interface to support an operator performing search and rescue missions. We performed the evaluation in a testbed that includes simulated hazards and rubble, and was based on a design developed by NIST for the RoboCup and AAAI Search and Rescue Competitions to represent an urban post-disaster environment. In these competitions, robot teams are allowed twenty minutes to locate and map simulated victims of a hypothetical disaster and are scored based on numbers of victims found, number and severity of unintentional contacts with victims or other objects in the testbed, and accuracy of the map showing victims' locations. Our evaluation requirements were influenced by the scoring rules, basic principles of human-computer interaction, intelligent systems qualitative criteria proposed by Messina, and evaluation issues articulated by Scholtz. The resulting evaluation yielded specific recommendations for improving the interface.

KEYWORDS: *interface evaluation, human-robot interaction, robot teams, search-and-rescue*

1. INTRODUCTION

We are concerned with the following problem: How can a human effectively control a team of robots in an urban-search-and-rescue (USAR) environment? Such a situation demands a well-designed human-robot interface that respects the cognitive and perceptual strengths and limitations of the human. Consequently, robot interface designers must be cognizant of what constitutes a "well-designed" interface.

This work was motivated by two factors. First, the MITRE robotics team plans to compete in the 2003 RoboCup Rescue Competition. To enhance our team's performance, we needed to understand the usability of the MITRE human-robot interaction (HRI) design and how it could be improved prior to the competition. Second, we wanted to demonstrate the feasibility of applying human-computer interaction (HCI) techniques to evaluate HRI, including developing usability requirements to establish a baseline usability level that can be referenced in further evaluations. (Usability requirements differ from functional requirements in that they take into account the performance of both the user and the system, although the focus is on the system, since that is what can be changed.) Scholtz [11] maintains that techniques from HCI can be adapted for use in HRI evaluation as long as they take into account the complex, dynamic, and autonomous nature of robots. We used time-tested HCI evaluation techniques that were developed as a result of decades of empirical research to recommend interface improvements and provide independent validation of MITRE's basic design approach. Researchers in the robotics field often face pressure to perform demonstrations that limits the opportunity to validate their approaches [6]; we hope to begin the process of breaking this trend by providing an example of applying usability evaluation techniques to a robotic interface.

We performed two types of evaluations using HCI techniques, one of which is reported on in this paper (the other is described in [3]). Prior to performing the evaluations, we defined goals and usability requirements so that we could better understand when improvements need to be made versus when the interface is "usable enough." We compared our findings with the requirements to determine the parts of the interface most in need of improvement.

Section 2 of this paper describes the MITRE robot's hardware and software, whereas Section 3 contains a description of the human-robot interface. Section 4 discusses previous work related to the evaluation of HRI, while Section 5 focuses on the evaluation. Section 6 provides results, followed by a discussion in Section 7.

2. OVERVIEW OF ROBOTIC COMPONENTS

The MITRE robot system is based on an ActivMedia Pioneer 2-AT robot platform. All interaction with the robot is mediated through an onboard computer running RedHat Linux with a 400MHz AMD K6-2 CPU and 256MB of RAM. The robot's onboard computer maintains an 802.11b wireless link to base station computers.

The robot's movement is generated through a four-wheel drive skid-steer system. To turn, the wheels on one side turn more quickly than those on the other. The faster-moving wheels travel farther than the slower pair, and all four wheels skid sideways to effect the turn.

The robot has a variety of sensors to provide information about its internal state and its surroundings. A ring of Polaroid sonar range finders returns information about the nearest detectable surface in each of sixteen directions. A SICK laser range finder returns very accurate ranges to diffusely reflective surfaces in a 180-degree swath in front of the robot. Finally, there is a video camera that can be panned, tilted, and zoomed as necessary. Pyrosensors (heat detectors) were added to the robot after the evaluation took place.

The robot software is written primarily in C++ and makes extensive use of an application programming interface (API) called ARIA (ActivMedia Robotics Interface for Applications) provided by the manufacturer of the robot.

Several autonomous behaviors help enable supervisory control of the robot team. Behaviors that enable exploration without requiring continuous monitoring include:

- *Wander* (randomly explore a region at a constant velocity, avoiding obstacles)
- *Go to* position (x, y), with destination selected by an operator gesture on the map display
- *Seek* certain types of regions (wander, but approach any region detected with certain defined characteristics)
- *Seek* with pyrosensor or video motion detection (implementation in progress)
- *Return* to home base (implementation in progress)

Additionally, MITRE roboticists have developed several behaviors that respond to direct tele-operation command by an operator. These commands specify robot velocity (forward-backward and turning), as well as pan-tilt-zoom controls for the onboard camera.

3. INTERFACE DESIGN

The MITRE command console interface includes several features that enable the operator to monitor and control up to three robots simultaneously. A single map pane indicates the locations of all robots and displays fused sensor data from all robots in a single world representation. The map contains multiple layers, which can be displayed individually or simultaneously:

- *Obstacle layer:* Represents the probability that a location is occupied by an obstacle to the robots. Data comes from robot bumpers, wheel encoders, sonar, laser range finder, and operator input.
- *Victim layer:* Represents the probability that a region contains a victim. Data comes from pyrosensor and motion detection algorithms, as well as operator input.
- *Explore/avoid layer:* Guides the operation of autonomous behaviors by representing regions that robots should explore or avoid. Data comes from operator input.

For each robot in the team (up to a current maximum of three robots), the interface provides a pane that includes a display of status messages, command history, and color video output, as well as a STOP control. Through status messages, the robots can alert the user when they need assistance or confirmation of a possible victim.

A depiction of one of the interface screens can be seen in Figure 1.

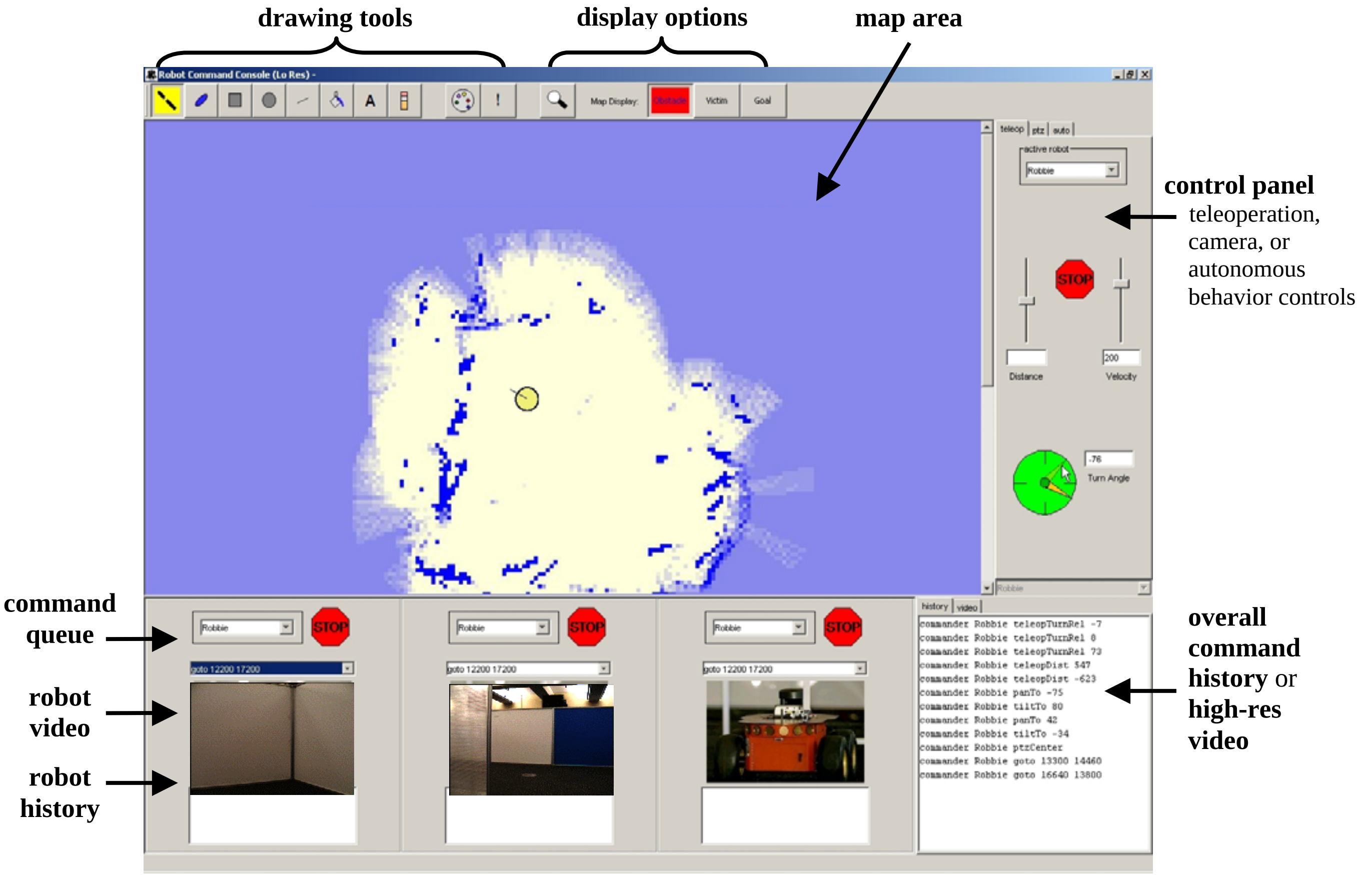
A final interface feature that supports supervisory control is the ability to queue autonomous commands. This capability enables the operator to give a sequence of commands to one robot – such as a series of "Go to" waypoints, followed by a "Pyrosensor seek" – then attend to other robots for a longer period of time while the first robot carries out the sequence autonomously.

4. RELATED HRI EVALUATION WORK

We needed a way of determining whether our interface approach would support the goal of being able to efficiently find victims in a search-and-rescue environment.

Evaluation of HRI can be viewed through the lenses of intelligent systems and HCI. Before any interface (robotic or otherwise) can be evaluated, it is necessary to understand the operators' relevant skills and mental models and develop evaluation criteria with those users in mind. There is no single, generally accepted set of evaluation criteria for HRI.

Messina et al. [8] proposed criteria as part of the intelligent systems literature. Messina's criteria are qualitative and apply to the performance of the robot only, as opposed to the robot and the operator(s) acting as a cooperating system. An example criterion from Messina is: "The system … ought to have the capability to interpret incomplete commands, understand higher level, more abstract commands, and to supplement the given command with additional information that helps to generate more specific plans internally" [8].



```
commander Robbie teleopTurnRel -7
commander Robbie teleopTurnRel 0
commander Robbie teleopTurnRel 73
commander Robbie teleopDist 547
commander Robbie teleopDist -623
commander Robbie panTo -75
commander Robbie tiltTo 80
commander Robbie panTo 42
commander Robbie tiltTo -34
commander Robbie ptzCenter
commander Robbie goto 13300 14460
commander Robbie goto 16640 13800
```

Figure 1. Command Console

In contrast, an example of a qualitative criterion that addresses human-robot performance might be: "The HRI shall facilitate the human's use of commands and his or her assessment of whether the robot has correctly interpreted and executed the commands." The latter, but not the former, criterion is an example of a usability requirement because to meet this requirement the interface must be tailored to the needs of the intended users.

Scholtz [11] proposes six evaluation "issues" for intelligent systems. Scholtz raises these issues "to determine what information the user needs to understand what the intelligent system is doing and when intervention is necessary, and what information is needed to make any intervention as effective as possible." Examples of Scholtz' issues are "Is the interaction language efficient for both the human and the intelligent system?" and "Are interactions handled efficiently and effectively – both from the user and the system perspective?"

When we defined evaluation criteria (stated as goals and usability requirements), they were informed by both Messina's criteria and Scholtz' issues, but were tailored for the specifics of the tasks expected to be performed by the MITRE robots and their operators.

Our evaluation took the form of a usability test. Typical users are asked to perform typical tasks, usually while "thinking aloud" [5]. "Thinking aloud" simply means the evaluation subject says whatever he/she is thinking as he/she uses the interface. HCI experts identify the components of the user interaction that cause problems for a majority of the users. We made use of the thinking aloud technique to help identify specific components of the MITRE robot interface that were difficult for users to operate.

There have been few studies of HRI that have made use of usability testing (or other HCI evaluation techniques). Three examples of studies that used versions of the usability test (although not the "classic" HCI approach we took) were performed by Casper [1], who evaluated search-and-rescue robot use at the World Trade Center disaster; Yanco [12], who evaluated a robotic wheelchair system; and Draper et al. [2], who tested a robot that re-arms military tactical fighters.

5. EVALUATION

Much research has gone into examining the effectiveness of usability testing (e.g., [7]). Usability testing was originally developed as a method in which large numbers of test subjects were given the same tasks, so that the results could be stated with a high degree of statistical confidence. As software development cycle times decreased and there was no longer sufficient time to perform an evaluation with many subjects, the necessity of attaining statistical significance was examined more closely. Researchers such as Jakob Nielsen determined that "even tests that are not statistically significant are well worth doing since they will improve the quality of decisions substantially" [9]. Nielsen and Landauer [10] developed a mathematical model to analyze the number of usability problems discovered versus dollars spent to discover them. They determined that "maximum benefit-cost ratio is achieved when using between three and five subjects" [9]. So, for example, one might expect to find approximately 80% of the problems using x subjects (where x = 3 to 5), rather than finding 98% of the problems (for example) using 10x subjects.

The key in choosing evaluation subjects, however, is to pick users who are truly representative of the user group for whom the interface is being optimized. In this case, the MITRE team wished to optimize the interface for the people who would be operating the robots at the 2003 RoboCup Rescue Competition. A group of four people are slated to operate the robots in the competition; we performed the usability test with three of these people.

The users' overarching goal was that they should be able to operate a robot or robot team efficiently. We assumed that means: users should have sufficient knowledge of the robot's state, be able to dynamically redirect the robot's behavior, control the robot with minimal memorization, and make a minimum number of errors. Specific requirements were derived that characterize the ability of the system to support the users in meeting their goals. The requirements are listed below in Table 1, along with the results showing which of these requirements were met.

5.1 Setup

The MITRE robot interface was set up on a laptop computer. One subject at a time was seated in front of the interface in an office approximately 20 meters from the arena as the crow flies; they were not able to see or hear the robots. The subjects' placement was a deliberate attempt to replicate the visual and aural isolation they will experience at the competition and to control for the communication lag likely at the competition. Commands were sent to the robot via the 802.11 network.

The MITRE robotics lab in McLean, VA contains a replication of a portion of the NIST standardized search-and-rescue arena (the "yellow," or least-challenging, portion); a single robot was placed in this arena during the usability test, at an initial location unknown to the subjects. We controlled the initial position of the robot and placement of the victims, and added clutter to the arena (papers, chairs, boxes, and a suitcase). We did not modify the interior or exterior walls of the arena, but the interior terrain was sufficiently altered from what users were used to seeing when viewed through the robot's sensors to provide a challenging environment.

Figure 2 shows a "bird's eye" view of the arena, where each square of the grid represents sixteen square feet. Each solid line represents a white, clear or blue, four-foot tall panel. Dashed light blue and dotted grey lines represent foggy or screen panels, respectively (also four-feet tall). The oval labeled "S" indicates the robot starting position, and the ovals labeled "V" show the victims' positions.

5.2 Conduct: Evaluators' Activities

Evaluator 1 was seated with the subject in the office. Evaluators 2 and 3 were in the lab with the robot. A phone line was left open between the two locations for coordination and timing purposes. The lab side was muted, though, so the subject could not receive any auditory cues.

The evaluators made a special note to look for cases in which the interface misled the subjects into operating the robot incorrectly (which we call "human errors"), and cases where the interface caused the subjects to express frustration or confusion. Also, the evaluators noted the following types of incidents:

- "Bumps": the robot made unintended contact with another object or victim
- "Command misinterpretation": the robot incorrectly executed a command
- "Total failure": a malfunction that required resetting one or more parts of the system

5.3 Conduct: Subjects' Activities

Subjects were asked to perform the following seven tasks during a 20-minute period, the amount of time they will have during the competition.

1. Put the robot in *teleoperation* mode. Drive the robot into the nearest room.
2. Give the robot a *Wander* command. Wait three minutes as it explores. If necessary, take control of the robot prior to the end of the three-minute time period if it appears likely to collide with an object.
3. Put the robot in *teleoperation* mode. Spend five minutes searching for victims.
4. Give the robot a *Go to* command. Did it go where it was supposed to go?
5. Give the robot a *Seek* command. Did it go where it was supposed to go?
6. Give the robot a *Wander* command. Have it *stop*. Did it stop immediately?
7. In the remaining time, try to find as many victims as possible.

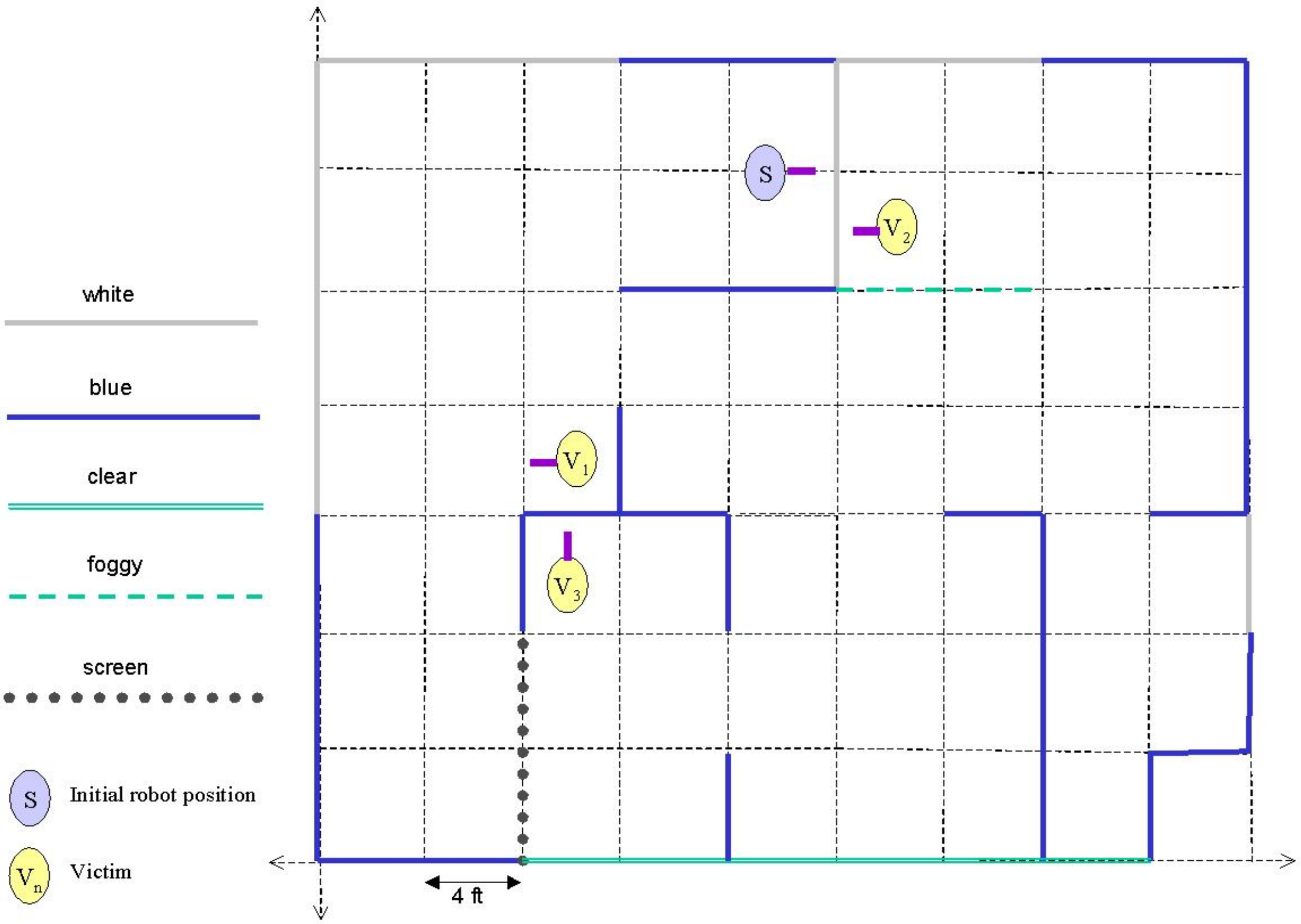


Figure 2. Arena Setup

6. RESULTS

Results are stated in Table 1 in terms of how many subjects were enabled by the interface to attain the required level of performance. (Note that the focus of each of the requirements is on the interface and not on the subject.) The requirements were driven by the objective of performing well at the RoboCup Rescue Competition and thus were heavily influenced by the scoring that will be used at the competition. Also, one of the requirements was influenced by [13], which found that the robot teams that covered more of the arena tended to find more victims and score higher in the competition.

Several aspects of the HRI design work very well. The interface does not require extensive manipulation to see all the relevant information, provided that one is using three or fewer robots simultaneously. The "STOP" buttons are highly visible and available in all modes of operation. The data fusion implemented in the map overlays is promising, as long as the response time can be decreased so that map updates do not significantly lag behind the robot's progress. In fact, the subject who successfully covered the largest fraction of the arena did so by relying almost solely on the map data (he turned on the video display late in the run after being prompted by an evaluator).

The three subjects found a total of only one simulated victim (which was spotted through a transparent panel). To find more victims and generally improve control of the robots in a search-and-rescue scenario, we feel the following modifications are critical:

- more timely map updates
- faster and more easily perceived feedback on acknowledgment and execution of commands
- additional context information regarding the robot's position
- more feedback concerning when a robot has made (or, better yet, is about to make) an error

Note that these recommended modifications all pertain to improving an operator's real-time awareness of a robot's position, immediate environment, and activities. (See [4] for a discussion of awareness in human-robot interactions.)

Table 1: Requirements and Results

Requirement	Results	Comments
The interface shall support users in identifying victims when the robot is within one meter of a victim	1 out of 3	One user identified one victim. That victim was identified via the video camera looking through a screen. The other users did not identify victims.
The interface shall support users in operating the robots such that they are aware whether a command has been executed correctly or incorrectly.	0 out of 3	All subjects had difficulty assessing whether the robot had responded correctly to their commands.
When in teleoperation mode, the interface shall support users in operating the robot such that only five bumps are incurred by each operator during a 20-min. search-and-rescue run in the arena.	3 out of 3	Although this requirement was met in all cases, note that there were 2 instances of significant bumps resulting in 2 total failures. Fewer bumps translate into higher scores.
The interface shall support users in operating the robots in such a manner that the robots can cover at least 50% of the arena within 20 minutes.	1 out of 3	The user who got the most coverage of the arena (50%) did so without consulting the video window. The other two users covered approximately 30% and 40% of the arena.
The interface shall be designed such that users will make a maximum of three human errors per 20-minute period.	3 out of 3	It was not obvious to subjects how to access the large video display (leading to errors) and one subject did not always use the distance slider correctly (he thought it was calibrated to different units), but no subject made more than 3 errors in 20 minutes.
The interface shall be designed such that no more than one total failure will occur during a 20-minute search-and-rescue run.	0 out of 3	Failure to meet this requirement was a result of the immaturity of the interface and robot software implementations at the time of evaluation.
The interface shall provide a means for users to interrupt the robot's current tasking and redirect the robot without having to wait for the robot to complete the original tasking.	2 out of 3	One subject encountered a bug after he attempted to issue a command that was not yet implemented; the robot waited indefinitely for the command to complete and ignored further tasking.
The interface shall support users in taking control of a robot that had been operating autonomously.	3 out of 3	The subjects were able to re-task the robot successfully after being directed to do so by the evaluators.
The interface shall support users in operating the robots such that they are aware that the robot has acknowledged commands.	0 out of 3	The interface has a command queue but only one subject consulted it, and did so infrequently.
The interface shall support users in understanding the robot's state such that users know when immediate intervention or corrective action is needed.	0 out of 3	In the cases of the two total failures, the subjects received no indication that they had knocked down walls in the arena until they were informed by the evaluators.
The interface shall be designed such that users do not have to remember information pertinent to controlling or monitoring the robot; the information shall be presented via the interface.	2 out of 3	The interface did not provide clues as memory joggers for how to access video; one of the subjects did not access video until prompted by evaluators late in the session.

7. DISCUSSION AND FUTURE WORK

Usability requirements are pertinent to the case of human-robot interaction because successful operation of a semi-autonomous robot requires a kind of partnership between the human and the robot: the robot must provide sufficient information to the human via the interface to enable him or her to make quality command and control decisions, and the operator must maintain sufficient awareness of the robots' activities to provide reasonable direction to the robots. Having these requirements focused our recommendations on improvements that will be likely to positively affect performance at the 2003 RoboCup Rescue Competition.

We plan to perform future HRI evaluations after the aforementioned improvements have been made, repeating the single-user, single-robot tasks outlined in this paper. Further evaluations may involve one user directing multiple robots simultaneously, and different user groups (e.g., search-and-rescue workers) under both single- and multiple-robot conditions. We will use the same requirements as criteria in further evaluations to determine whether there has been an improvement. We plan to examine MITRE's scores at the RoboCup Rescue Competition to see how well adherence to the requirements predicted success at the competition, and to evolve the requirements based on what we learn.

8. REFERENCES

[1] Casper, J. Human-Robot Interactions During the Robot-Assisted Urban Search and Rescue Response at the World Trade Center. MS Thesis, University of South Florida Department of Computer Science and Engineering, 2002.

[2] Draper, J. V., F. G. Pin, J.C. Rowe, and J. F. Jansen. "Next generation munitions handler: human-machine interface and preliminary performance evaluation." In Proceedings of the 8th International Topical Meeting on Robotics and Remote Systems, Pittsburgh, PA, 1999.

[3] Drury, J., L. D. Riek, A. D. Christiansen, Z. T. Eyler-Walker, A. J. Maggi, and D. B. Smith. "Command and Control of Robot Teams." In Proceedings of the 2003 Association of Unmanned Vehicles International (AUVSI) Conference, Baltimore, MD, 2003.

[4] Drury, J. L., J. Scholtz, and H. A. Yanco. "Awareness in Human-Robot Interactions." Accepted for publication in the *Proceedings of the 2003 International Conference on Systems, Man, and Cybernetics*, Washington D. C., 2003.

[5] Ericsson, K. A. and H. A. Simon. "Verbal reports as data." *Psychological Review*, vol. 87, pp. 215 – 251, 1980.

[6] Harrigan, R. W. "Is it a wave or is it a particle? The synergy of theory and experimentation." Workshop on Validation of Public Sector Robotic Systems: Moving from Demos to Experiments, 2002 IEEE International Conference on Robotics and Automation, Crystal City, Virginia, 2002.

[7] Mantei, M. M. and T. J. Teorey. "Cost/Benefit Analysis for Incorporating Human Factors in the Software Lifecycle." *Communications of the ACM*, vol. 31 number 4, 1988.

[8] Messina, E., J. Evans and J. Albus. "Evaluating knowledge and representation for intelligent control." In Proceedings of the 2001 Performance Metrics for Intelligent Systems (PerMIS) Workshop, in association with IEEE CCA and ISIC, Mexico City, Mexico, 2001.

[9] Nielsen, J. "Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier." In *Cost-Justifying Usability*, R. G. Bias and D. J. Mayhew, editors, 1994.

[10] Nielsen, J. and T. K. Landauer. "A mathematical model of the finding of usability problems." In Proceedings of the ACM INTERCHI'93 Conference, Amsterdam, The Netherlands, 1993.

[11] Scholtz, J. "Evaluation methods for human-system performance of intelligent systems." In Proceedings of the 2002 Performance Metrics for Intelligent Systems (PerMIS) Workshop, Gaithersburg, MD, 2002.

[12] Yanco, H. A. "Shared User-Computer Control of a Robotic Wheelchair System." Ph.D. Thesis, Massachusetts Institute of Technology Department of Electrical Engineering and Computer Science, 2000.

[13] Yanco, H. A., J. L. Drury and J. Scholtz. "A Study of Human-Robot Interaction at the AAAI-2002 Rescue Competition." Accepted for publication in the *Journal of Human-Computer Interaction* in 2004.