# Evaluation of a Human-Robot Interface: Development of a Situational Awareness Methodology

Jean Scholtz, Brian Antonishek, Jeff Young
*National Institute of Standards and Technology*
*{jean.scholtz, brian.antonishek, jeff.young}@nist.gov*

## Abstract

*This paper outlines a methodology to evaluate supervisory user interfaces for robotic vehicles based on an assessment of situational awareness. The results of an initial experiment are discussed. The evaluation method will be validated in a future experiment that will also result in a benchmarked user interface.*

## 1. Introduction

Our research agenda is to develop guidelines for user interfaces in human-robot interaction. In order to accomplish this we need to first understand what information and interactions are needed for effective and efficient human-robot interaction.

The term "robot" must first be defined as a number of interpretations are possible. Knowledge robots (commonly referred to as "bots") are computer software robots that continuously run and respond automatically to a user's activity. The Oxford English Dictionary defines "robot" as an "apparently human automaton, intelligent but impersonal machine." According to the Australian Robotics and Automation Association (ARAA), there *is* no standard definition. The ARAA suggests that a robot has "three essential characteristics:"

- It possesses some form of mobility
- It can be programmed to accomplish a large variety of tasks
- After being programmed, it operates automatically.

The Collins English Dictionary defines a robot as "any automated machine programmed to perform specific mechanical functions in the manner of a man." The Cambridge International Dictionary of English defines robot as a machine used to perform jobs automatically which is programmed and controlled by a computer.

The dictionary entry also notes that some types of robots can walk and talk, but they cannot think like humans. Murphy [10] defines an intelligent robot as a mechanical creature that can function autonomously. While robots have computers as an integral component, we normally do not think of computers as being mobile, interacting with and changing aspects of the physical world. We are not concerned with software robots in our research. Our emphasis is on mobile mechanical devices with some degree of autonomy.

One question that might be asked is why a different focus is needed for human-robot interaction (HRI) and how that differs from human-computer interaction (HCI)? Scholtz [13] noted six dimensions in which HRI and HCI differ:

1. Different requirements based on interaction roles
2. Interaction of the platform in the physical world
3. Dynamic nature of the hardware
4. Environment in which the interactions occur
5. Number of platforms that the user is interacting with
6. Autonomous behavior of the platform

Fong [7] also notes that HRI differs from HCI because HRI is concerned with complex, dynamic controls systems operating with autonomy in real-world environments. Scholtz [14] defines the following roles:

- Supervisor
- Operator
- Mechanic
- Team mate or peer
- Bystander

While multiple roles may be played by one person interacting with the robot, a given robot may have multiple users interacting at any one time in various roles.

To investigate this area large systematically, we are considering smaller subdivisions. The basis for our subdivision is based on the definition of interaction roles and domains in which robots are utilized. We have taken several methodological approaches to investigating this area. We have performed empirical studies in a laboratory setting [15] as well as studying critical incidents in real-world settings [3,16,17]. We are currently looking at user interfaces for operators in search and rescue and autonomous driving domains. We are also investigating the difficulty of creating models of interactions in the bystander role.

The work presented in this paper is focused on developing an evaluation methodology for the supervisory role in robotic vehicles for on-road driving. Our hypothesis is that supervisors of robotic vehicles need the ability to monitor multiple vehicles and need to be able to quickly determine which, if any, vehicles demand closer attention. We have developed an evaluation methodology based on situation awareness assessment to assess user interfaces designed to support the supervisory role.

## 2. Situation Awareness Assessment

Operator interfaces in such areas as air traffic control, aircraft, power plants, and manufacturing have been evaluated with regard to the amount of situation awareness provided. Endsley [4] identifies three levels of situation awareness: perception, comprehension, and projection. Perception is the basic level of situation awareness (SA level 1). This level of awareness is achieved if operators are able to perceive in the user interface the information that is needed to do their job. The next level is comprehension (SA level 2). Not only must the information be perceived, it must be combined with other information and interpreted correctly. The third level (SA level 3) is projection or the ability to predict what will happen next based on the current situation. Time is an element of situation awareness as well. Situations are dynamic. An operator may have situation awareness for a short period of time but then be flooded by information and not have situation awareness as events unfold rapidly. Operators must also be able to predict how much time they will have before an event occurs so that they can judge the interactions necessary to react.

Several methods have been used to assess the situation awareness delivered by the user interface [12]. A performance based method looks at the actions taken by the operator and determines the appropriateness or correctness of the actions. The problem with using performance based measures is that an operator can have perfect situation awareness and still perform inappropriate actions due to suboptimal decision making. Failures can also be attributed to problems with the user interface. Environmental factors could also complicate this process. As many responses are appropriate in different situations, it is difficult to measure situation awareness unambiguously. Knowledge-based measures and verbalization measures of situation awareness are also used. Knowledge-based measures isolate components of situation awareness and assess these separately. This type of methodology is potentially intrusive into the operator's task and in general, is not used in actual tasks but is assessed in simulated environments. Verbalization methods provide insight about the importance of information and the process that the operator uses. Again, this methodology can be intrusive to the operator's task. Both of these methods can identify declarative knowledge better than procedural knowledge. Performance-based methods give better information about procedural knowledge. Subjective measures of situation awareness are also used [8]. Participants use a linear scale and assign a numerical value to their situation awareness. Observers may also assign these numerical scales based on predetermined categories that have been identified as critical to a certain process [1].

Our work is modeled after the Situational Awareness Global Assessment Technique (SAGAT) developed by Endsley [5]. This evaluation methodology uses expert knowledge to develop questions that assess the users' awareness of a particular situation. A simulation is used and the user is stopped during the simulation and given a quick series of questions to answer. These questions assess the three levels of situational awareness. After answering these questions, the users are returned to the simulation. Endsley found little interruption when the time to answer assessment questions is limited to several minutes.

## 3. Experimental design

Our research has two objectives. The first is to develop an evaluation methodology for user interfaces for the supervisory role in human-robot interaction. In particular, we have chosen the domain of on-road driving as our example. The second objective is to conduct the necessary empirical studies to produce a user interface that can be used as a benchmark or baseline in this area.

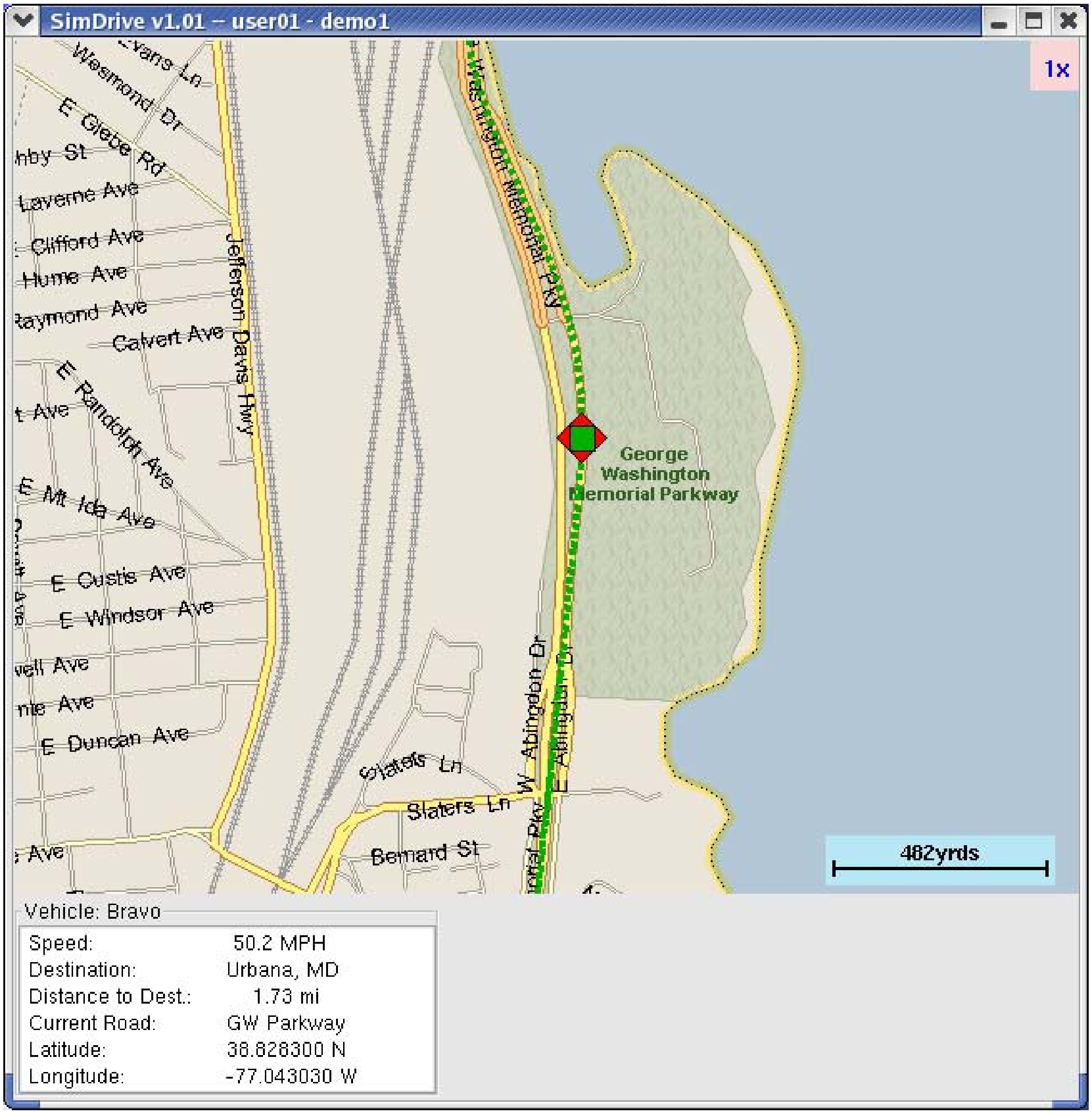



Figure 1: Overall view of the supervisory user interface for on-road driving

### 3.1. Description of the user interface

We designed the user interface with two objectives in mind: to make it easy for the user to understand when a vehicle was in trouble and might have to be more closely monitored and to produce a design that could scale to support a number of vehicles. Our long term goal is to develop a baseline user interface for monitoring multiple vehicles. We used a map background for the user interface for our on-road driving vehicle. In this experiment we are only concerned with monitoring one vehicle but we were aware during our design that the user interface had to scale to support a number of vehicles.

We classified information that a supervisor would need to know into three types: vehicle information, environment information, and route information. Vehicle information consists of speed of the vehicle, fuel level of the vehicle, status of any sensors on the vehicle, the current position of the vehicle including the roadway it is on, the lane position, and the driving behavior currently being executed (e.g. turning left, stopped at traffic control). Environment information is the traffic around the vehicle, the posted speed limit, the road conditions including any obstacles on the roadway, and the status of any traffic control signals or any road signs currently in view. Route information includes the distance to the final destination, the speed of the vehicle, the destination, and the current roadway name or number. Figure 1 shows the overall view of our user interface.

The route information is displayed in the bottom portion of the screen. When a number of vehicles are monitored, this area will have tabs labeled with vehicle identification. Route information for a vehicle will come to the front by selecting that tab or by selecting the vehicle icon directly on the map display.

We used a two part icon to indicate the status of the vehicle and environment. We used both shape and color [9] to indicate whether the status was normal, cautionary or hazardous. As shown in figure 2 the square is the shape used for a normal condition, the circle for caution, and the diamond for trouble. The outer shape and color represents the environment status and the inner shape and color represents the vehicle status.

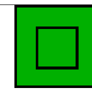
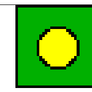
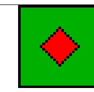
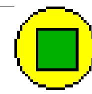
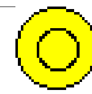
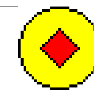
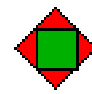
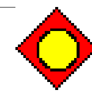
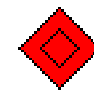
| *STATUS ICONS* | *Vehicle Normal* | *Vehicle Caution* | *Vehicle Trouble* |
|---|---|---|---|
| *Environment Normal* |  |  |  |
| *Environment Caution* |  |  |  |
| *Environment Trouble* |  |  |  |

Figure 2: Icons used to indicate status of the vehicle and the environment

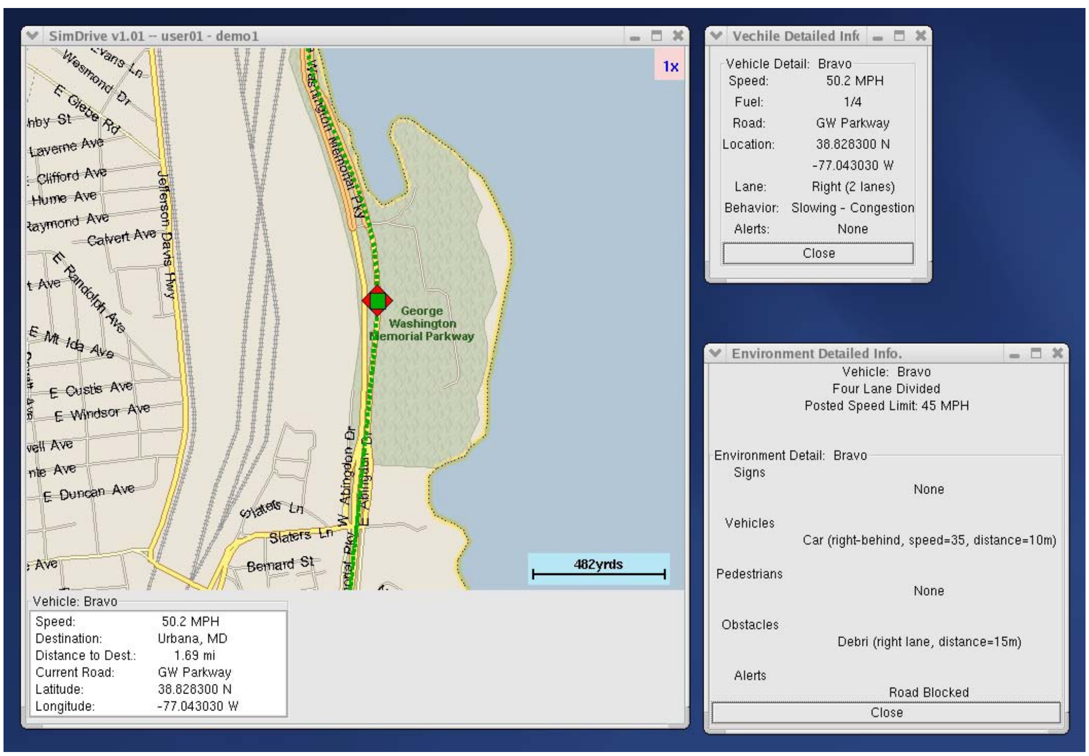



Figure 3: The supervisory interface with the vehicle and environment status windows

The triggers for cautionary and dangerous conditions are settings that the supervisor would most likely specify. Cautionary displays might indicate such things as a number of vehicles in lanes beside the vehicle, traffic signals being approached, or pedestrians nearby. Hazardous conditions might include obstacles in the roadway, vehicles pulling in front of this vehicle, or unsafe road condition. While the icon shows the overall status, the supervisor can obtain specific information about the vehicle and the environment by pulling up a window with this information. Currently this is accessed by using control keys. When multiple vehicles are monitored, we envision the supervisor being able to select the vehicle icon directly on the map or make a selection from a panel containing the names of all the vehicles. Figure 3 shows the user interface with the vehicle and environment windows open.

## 3.2. Assessment methodology

We employed the SAGAT style method of assessing situational awareness. Participants in our experiment were asked to monitor a simulation that showed the vehicle navigating through the environment. The vehicle, environment and route information were continually updated. At predetermined times, the simulation froze and the participant was asked to turn to a data collection computer and was asked to indicate the current situation of the vehicle in the simulation. We developed three types of questions, one type for assessing each level of situational awareness. A Visual Basic user interface was developed to collect this data. Figure 4 shows one of the data collection screens used in the experiment. The questions for SA level 1 and SA level 2 were always the same, while the question for SA level 3 differed depending on the scenario.

The SAGAT methodology uses experts to develop the information that is used to assess situational awareness. In our case, we used a computerized driving training program to determine types of information that should be used to assess the different types of situational awareness (Driver-ZED[1]).

To assess SA level 1 we asked participants the status of the vehicle, the environment, and route information, namely the distance to the destination. We designed the user interface to clearly display these indicators. The icons displayed the vehicle and environment condition and the route information was displayed at the bottom of the map. We asked participants to indicate risks at the point in time when the simulation was frozen by selecting from a list of 12 possibilities. The number of risks varied depending on the scenario, varying from 1 to 3, but the list of choices was always the same. This information could be obtained from the vehicle and environment status windows. The user has to perceive the information in the display and understand that this is a potential danger. Therefore, assessing the participants' awareness of potential risks is a level 2 situation awareness.

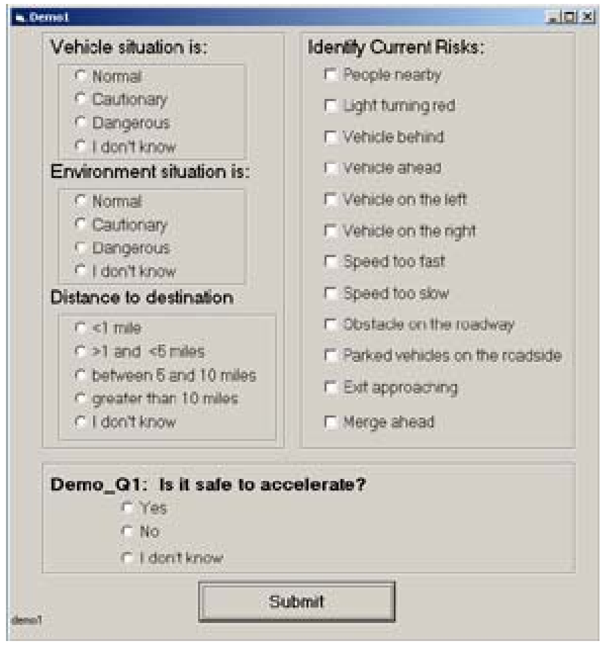



Figure 4: The data collection screen used in our experiment

To assess level 3 awareness, we asked the participants to use their awareness of the vehicle, environment, and route to answer such questions as "is it safe to make a left turn now?" Participants could choose 'yes', 'no', or 'I don't know' as answers. We included the 'I don't know' category to discourage guessing.

## 3.3 Simulation details

The 'SimDrive' simulator program was written starting with 'GPS-Drive' as a base to utilize its routines for displaying maps and for working with GPS data points. GPS-Drive is GPS navigation system which runs on Linux and is written in C using the GTK+ graphics library. GPS-Drive is written under GPL (GNU General Public License, http://www.gnu.org/copyleft/gpl.html). GPS-Drive binaries and source code are available at http://gpsdrive.kraftvoll.at/index.shtml.

Developing the simulator involved adding routines for viewing simulated driving scenarios, creating a simulation file editor to add details to those scenarios, and customizing the user interface to be able

---

[1] Use of this product does not constitute endorsement by the National Institute of Standards and Technology.

to present all of the new information to the participant. These additions were also written in C using the GTK+ graphics library.

When a participant starts a session his corresponding user file is selected. This user file simply contains the pre-determined random order in which the simulation scenarios will be presented to that particular participant. This order is also the same order in which the SA 3 questions are asked, thus assuring that the correct SA 3 question is asked for each scenario that is presented.

All the information presented to a participant during a scenario is contained in that scenario's simulation data file. This simulation file is comprised of a list of GPS coordinates that represent the path a vehicle will follow during that scenario. Along with each GPS data point are the current values of all the data fields that are presented during the scenario such as vehicle status, road description, etc.

To simulate realistic traffic conditions that a supervisor might encounter, traffic speed and position data was collected using local roads and the corresponding traffic conditions. A hand-held GPS unit was used to collect current position data during routine trips within a 40 mile radius at various times of day and over varied road types (interstate, highway, city, and neighborhood). GPS data points were stored once every second and are presented to the participant, during the simulation, at the same rate.

The collected GPS data points were then downloaded from the hand-held unit and translated for use with the simulator. Then using the simulation file editor, appropriate scenario related information details were added to these GPS data points. The editor allowed us to step through a simulation file and visualize the vehicle position in the simulator and to set the desired vehicle and environment values at that position of the scenario.

### 3.4 Experiment details

Participants for the experiment were recruited from colleagues at work. Some had experience with robots, while others did not. All of the participants held a valid drivers license.

We developed three scenarios for training the participants. We explained the purpose of the experiment and the different portions of the user interface to them. We then asked the participants to monitor the vehicle and to answer the situational awareness questions for the three training scenarios. We developed ten scenarios for the experiment. These varied from highway driving to urban driving and contained different types of potential hazards, including traffic, traffic signals, pedestrians, and obstacles on the road. The scenarios were presented in a random order to the participants to eliminate possible biases based on learning. Each scenario lasted between 1 and 2 minutes. After participants had completed this part of the experiment we gave them three longer scenarios (between 2 ½ and 3 ½ minutes) to monitor. After each of these scenarios we administered the NASA Task Load Index [11] workload questionnaire to determine the cognitive load imposed on the participants.

## 4. Results

We scored the situational awareness questions as follows. There were three SA level 1 questions; one each for the vehicle, environment, and route. Participants were given 1 point for each correct answer. Thus the total number of points possible for each scenario was 3. Figure 5 shows the Level 1 results for the 10 scenarios.

The results for SA 1 were good. The lowest scores were for scenarios 5 and 9, but the mean score was still 2.5 or over. Participants were able to clearly see the condition of the vehicle, the environment and the route. We do not understand why the scores for scenarios 5 and 9 were lower than the other scenarios. Half of the participants in scenario 9 got only 2 of the 3 answers correct. For scenario 5 3 participants got 2/3 correct while one participant was able to get only 1 answer correct. We are considering trying to run the scenarios again using a talk-aloud protocol [6] to better understand the reasons for participants' choices.

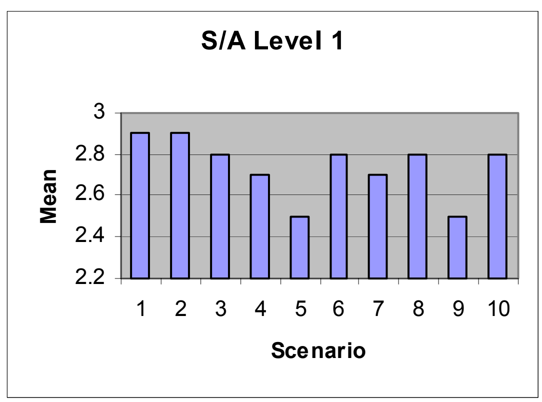



Figure 5: Results for Situational Awareness level 1

SA level 2 had a different number of possibilities for each scenario. We again gave the participants 1 point for each correct answer but subtracted a point for any additional risks they checked. Figure 6 shows the results for SA level 2. Note that scenarios 3,4 and 5 had a max score of 2, scenarios 6 and 9 had a max of 3, and all other scenarios had only 1 potential risk.

On average the results were reasonable but scenarios 1, 2, 8 and 10 were lower than the others. Scenario 2 (speed too fast condition) may have been problematic as participants needed to determine a variable in the environment box (posted speed) and a variable in the vehicle box (current vehicle speed) and also to recall initial directions about what constitutes excessive speed. Another problematic issue was the description of traffic near the vehicle. We described other traffic in several instances as 'right-ahead' or 'left-ahead'. We expected participants to indicate potential risks as both traffic to the right or left and traffic ahead. However, in a number of instances participants selected only one risk, either vehicle right or left or vehicle ahead. We need to ensure that our language or representation in future experiments is less ambiguous.

There was only one question for SA level 3. We gave participants one point for being correct. Figure 7 shows these results. Half of the scenarios were above 60% which is reasonable. We looked at the other scenarios to determine what the difficulty with those scenarios might have been. We offered participants an 'I don't know' choice but only 5 of the incorrect answers were "I don't know." The big issue is that in a number of instances the vehicle was traveling on a multilane road. We used text to represent the lane that the vehicle was in and the lanes that surrounding traffic was in. As with the assessment of level 2, this seems to be problematic. For example, in scenario 9 there is an obstacle in the road so we expected the answer to 'do you have a potential situation to deal with?' to be 'yes'. However, as the robotic vehicle was in the right lane and the debris was in the left lane, participants might have answered 'no'.

We had originally anticipated that participants would bring up the environment and vehicle status boxes only when the status was cautionary or dangerous. We found that participants brought these windows up at the beginning of the scenarios and arranged them so that they were visible for the entire scenario. Once we move to displaying multiple vehicles there will be too many windows for participants to do this. We had provided logging capabilities in the simulation software to track the times and durations that the status boxes are open. As these boxes were always open during this experiment, this information serves no purpose.

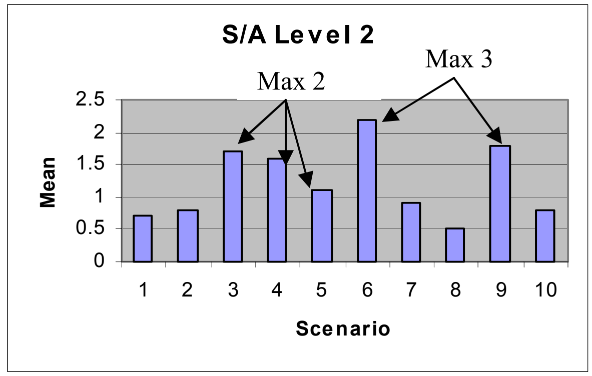



Figure 6: Results for Situational Awareness level 2

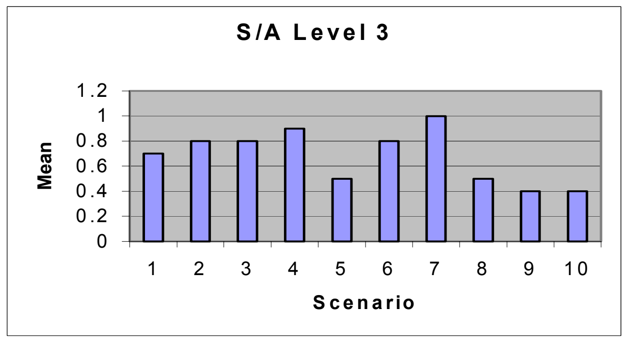



Figure 7: Results for Situational Awareness level 3

The NASA TLX results were computed for the three longer scenarios. These three scenarios were designed to reflect different levels of difficulties for the vehicles. These three scenarios lasted between 2 1/2 and 3 1/2 minutes. The first scenario had only one cautionary condition (speed over the limit) that lasted for 40 seconds. The second scenario had two periods of caution, one for the vehicle and one for the environment, and two periods of vehicle trouble. The third scenario had one vehicle caution period, one environment caution, and one environment trouble period. The environment caution lasted for over 100 seconds (severe weather).

We wanted to determine what the workload was on the participants using the given user interface for monitoring the vehicle status. As seen in Table 1, the workload for the three scenarios was relatively consistent for all three but there was a large range in the workload scores. Table 1 suggests that Long

Scenario (LS) 2 was slightly more difficult than LS 1 and 3 (because the sum of 2 is higher, meaning that people consistently rated it higher than LS 1 and 3).

Table 1. Descriptive Statistics for Weighted Workloads for Long Scenarios 1 thru 3

| | Range | Min | Max | Sum | Mean | Std. Dev. |
|---|---|---|---|---|---|---|
| S 1 | 67 | 10 | 77 | 351 | 35.10 | 24.30 |
| S 2 | 45 | 21 | 66 | 394 | 39.40 | 15.31 |
| S3 | 69 | 8 | 77 | 336 | 33.60 | 20.94 |

*N* = 10

Table 2 shows the means for the individual participants and shows how the workload measures vary across the participants.

Table 2. Weighted Workload means for individual participants across Long Scenarios 1 thru 3.

| Participant | Mean |
|---|---|
| 1 | 51.33 |
| 2 | 73.33 |
| 3 | 21.67 |
| 4 | 16.00 |
| 5 | 14.00 |
| 6 | 29.33 |
| 7 | 51.33 |
| 8 | 37.67 |
| 9 | 45.00 |
| 10 | 20.67 |

Table 3 shows the correlation of the workload measures for the three scenarios. Table 3 also provides evidence that participants found LS 2 more difficult. LS 2 is highly correlated with LS 1 and 3 despite having a higher sum, implying that all participants tended to score LS 2 higher than LS 1 and 3. The descriptions of LS 2 and LS 3 do not seem on the surface to show much difference. However, plotting these two scenarios over time does suggest that in LS 2 the environment caution overlaps a vehicle caution event and two vehicle danger events. This suggests that LS 2 is somewhat more difficult than LS 3 as suggested by the higher workload scores for monitoring this scenario.

Table 3: Correlation of Weighted Workloads for Long Scenarios 1 through 3

| | 1 | 2 | 3 |
|---|---|---|---|
| Long Scenario 1 | | | |
| Long Scenario 2 | .89* | | |
| Long Scenario 3 | .85* | .82* | |

*p < .01; *N* = 10

## 5. Discussion

We think that our assessment technology is viable but we need to refine our scenarios to ensure that participants can interpret them unambiguously. The results for SA 1, SA 2, and SA 3 lead us to conclude that our user interface is reasonable but needs improvement, especially for SA 2 and SA 3. We have ideas for future improvements, such as making the vehicle and environment status windows more graphical. That is, showing the vehicle in relationship to surrounding traffic graphically rather than relying on textual descriptions.

We also need to investigate workload more thoroughly. While the workload measures did tend to show that LS 2 was more difficult for the subjects we need to ensure that we have a more pronounced difference to adequately assess how our user interface functions for difficult scenarios.

We also had a wide range of workload scores across individuals. We need to determine if such individual differences are expected. We did show the scenarios in the same order, so all participants saw scenario 3 last. This may have also have had an effect on their workload measure. In addition, as this was not an interactive simulation, participants did not have to perform some sort of action. These are issues we need to address in the next set of experiments.

We have two next tasks. The first is to validate the sensitivity of our situation awareness assessment methodology. We intend to do this by improving the user interface and running a second experiment. This experiment will include refined assessment methods as we have already described. Once we are satisfied with the user interface benchmark for one vehicle, our second task will be modifying the improved user interface to handle multiple vehicles and to run an experiment to measure the situation awareness.

## 6. Conclusions

We are in the process of developing a methodology for assessing the situational awareness provided by user interfaces for monitoring on-road driving vehicles. While we have focused our research in this particular domain, we feel that this methodology could certainly be adapted to other domains where obtaining situational awareness is critical to the user. We have implemented the methodology and have designed and measured a supervisory user interface using this methodology. The user interface and the simulation program are available for others who would like to make improvements to the interface or to use the methodology to assess other user interfaces.

## 7. Acknowledgements

This work was funded by the DARPA MARS research program. Paul Hsiao of NIST designed the data collection user interface. Thanks to the many NIST summer students who participated in this experiment.

## 8. References

[1] Bell, H. and Lyon, D. Using Observer Ratings to Assess Situation Awareness (2000) in Mica R. Endsley and Daniel J. Garland (Eds.) *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates, Mahwah, New Jersey.129-146.

[2] Drivers-ZED® for Adult Drivers, AAA foundation for Traffic Safety, 1-800-305-SAFE.

[3] Drury, J, Scholtz, J, and Yanco, H. 2003. Awareness in Human-Robot Interactions, to be presented at 2003 IEEE International Conference on Systems, Man & Cybernetics.

[4] Endsley, M. 2000. Theoretical Underpinning of Situation Awareness: Critical Review (2000) in Mica R. Endsley and Daniel J. Garland (Eds.) *Situation Awareness Analysis and Measurement.* Lawrence Erlbaum Associates, Mahwah, New Jersey. 3-32.

[5] Endsley, M. R. 1988. Design and Evaluation for situation awareness enhancement. *In Proceedings of the Human Factors Society 32$^{nd}$ Annual Meeting,* vol. 1, 97-101. Santa Monica, CA: Human Factors Society.

[6] Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: The MIT Press.

[7] Fong, T., Thorpe, C. and Bauer, C. 2001. Collaboration, Dialogue, and Human-robot Interaction, *10$^{th}$ International Symposium of Robotics Research*, November, Lorne, Victoria, Australia.

[8] Jones, D. Subjective Measures of Situation Awareness (2000) in Mica R. Endsley and Daniel J. Garland (Eds.) *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates, Mahwah, New Jersey. 113-128.

[9] Mayhew, D. 1992. *Principles and Guidelines in Software User Interface Design,* Englewood Cliffs, NJ: Prentice Hall, 494.

[10] Murphy, R. 2000. *Introduction to AI ROBOTICS.* Cambridge, Massachusetts : MIT Press.

[11] NASA Task Load Index (TLX), Human Performance Research Group, NASA Ames Research Center, Moffett field, California, v.1.0.

[12] Pritchett, A. and Hansman, R. Use of Testable Responses for Performance-Based Measurement of Situation Awareness (2000) in Mica R. Endsley and Daniel J. Garland (Eds.) *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 189-209.

[13] Scholtz, J. 2002. Creating Synergistic CyberForces in Alan C. Schultz and Lynne E. Parker (eds.), Multi-Robot Systems: From Swarms to Intelligent Automata. Kluwer.

[14] Scholtz, J. 2003. Human-robot Interactions: Creating Synergistic Cyberforces. Hawaii International Conference on System Science, Jan. 2003.

[15] Scholtz, J. and Bahrami, S. 2003. Human-Robot Interaction: Development of an evaluation methodology for the Bystander Role to be presented at 2003 IEEE International Conference on Systems, Man & Cybernetics.

[16] Scholtz, J., Antonishek, B., and Young, J. 2003. Evaluation of Operator Interventions in Autonomous Off-road Driving, to be presented at PerMIS 2003.

[17] Yanco, H., Drury, J. and Scholtz, J. In Press. Beyond Usability Evaluation: Analysis of Human-Robot Interaction at a Major Robotics Competition. To appear in *Human-Computer Interaction*