Evaluation Methods for Human-System Performance of Intelligent Systems

Dr. Jean Scholtz
Visualization and Usability Group
Information Access Division
Information Technology Laboratory
National Institute of Standards and Technology

Abstract

Intelligent systems are becoming more and more of a reality but with the exception of very special purpose systems, completely autonomous systems are not yet the norm. In reality, we need to have humans who monitor the systems, intervening when necessary. As systems increase in intelligence, the goal for human-in-the-loop activities should not be to eliminate the human, but rather to create a human-system partnership with greater capabilities than the individual components. We currently view intelligent systems and the operators or supervisors of these systems as separate components and conduct evaluations in the same vein. For intelligent systems to become more useful and acceptable, we need to consider the "system" as a synergistic composition of software behaviors, possibly embedded in a physical component such as a robot, and the human interacting with this virtual or physical component. Our objective is to design this team interaction in such as way that the intelligence of the team is greater than the intelligence of any one of the parts.

Keywords: *Human-robot interaction, situational awareness, human-computer interaction, evaluation methodologies, intelligent systems.*.

1. Introduction

In our work we are concerned with intelligent systems embodied in hardware (robots). Human-robot interaction is fundamentally different from typical human-computer interaction (HCI) in several dimensions. [8] notes that HRI differs from HCI and Human-machine Interaction (HMI) because it concerns systems which have complex, dynamic control systems, exhibit autonomy and cognition, and which operate in changing, real-world environments. In addition, differences occur in the types of interactions (interaction roles), the physical nature of robots, the number of systems a user may be called to interaction with simultaneously, and the environment in which the interactions occur.

The interaction roles of supervisor, operator, and peer are defined in [17]. Upon further consideration we have subdivided two of these roles resulting in five different interaction roles: supervisor, operator, mechanic, teammate, and peer. The supervisory role involves monitoring the intelligent system and seeing that any interventions that are needed are handed off to the proper individual. We have subdivided the original operator role into an operator role and a mechanic role. An operator is needed to work "inside" the robot; adjusting various parameters in the robot's control mechanism to modify abnormal behavior; to change a given behavior to a more appropriate one; or to take over and tele-operate the robot. The mechanic interaction is undertaken when a human needs to adjust physical components of the robot, such as the camera or various mechanical mechanisms. The peer role has been divided into a teammate role and a bystander role. The teammate role implies the same relationship between humans and robots as it does in human-human interactions. Teammates of intelligent systems can interact at an "implementation level." The commands a teammate can give to a robot should not change the nature of the plan or mission but allows adjustments due to the dynamics of a particular situation. A bystander does not explicitly interact with a robot but needs some model of robot behavior as the bystander will be in the same physical space as the robot and needs to co-exist.

The second dimension is the physical nature of mobile robots. Robots need some awareness of the physical world in which they move. As robots move about in the real world, they build up a "world model" [2]. The robot's model needs to be conveyed to the human in order to understand decisions made by the robot as the model may not correspond exactly to reality due to the limitations of the robot's sensors and processing algorithms.

A third dimension is the dynamic nature of the robot platform and its effect on performance and capabilities. In typical human-computer interactions the assumption is that the computer is working and that behavior does not change over time. In assessing user interactions with the internet, we are starting to question this assumption as the workload and the time delays at any particular time can affect what the user does and how satisfied the user is with the experience. The fact that robot capabilities can change implies that functionality at time 1 may not be

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2002		**00-00-2002 to 00-00-2002**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Evaluation Methods for Human-System Performance of Intelligent Systems	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institute of Standards and Technology (NIST),Manufacturing Engineering Laboratory,100 Bureau Dr,Gaithersburg,MD,20899**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the 2002 Performance Metrics for Intelligent Systems Workshop (PerMIS ?02), Gaithersburg, MD on August 13-15, 2002

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

available at time 2 and this has to be factored into the human-robot interaction.

In typical human-computer interactions the cognitive state of the human has been largely ignored. The notion of affective computing [15] is just starting to appear in commercial products. Human computer interaction specialists in such domains as air traffic control [12], safety critical systems, and military systems have been concerned with these issues for some time and have attempted to designed user interfaces that are usable in these conditions. The current trend is to design interfaces that detect the status of the user and adjust accordingly. Consideration of environmental conditions and the effect on the users is a necessity in HRI.

In human-robot systems several of the interaction roles (supervisor, teammate, bystander) may need to interact with a number of systems simultaneously. These systems may be operating completely independently or they may be functioning as a team. Moreover, several types of interactions could be occurring at the same time; for example, a supervisor might be overseeing a robot which is also interacting with a teammate. Typical HCI considers 1:1 situations; one user to one computer. In HRI, we have the possibility of a 1:N situation where one person is interacting with N robots and N:1 where a number of people are interacting with the same robot.

The autonomous nature of intelligent systems has been a subject of discussion for some time in the human factors world [13]. This changes the user's role from being in control to monitoring and intervening when necessary. This introduces the concept of "being out-of-the-loop" and raises the issue of how to alert the user to an exception and how to bring the user up to speed to quickly and effectively intervene.

2. Approach

Traditional HCI takes a user-centered approach [14] while others in the automation field have taken a system-centric approach [3]. We are taking an information-centric view. It is necessary 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. This understanding is basic to the design of a user interface that is able to present the appropriate information to the user. Intervention requires one more component: a language for the user and the intelligent system to use in resolving the problem. The final part of the problem is actually carrying out the intervention, assessing the situation correctly and giving advice or performing a necessary action for the intelligent system.

We propose six different issues in evaluation that must be considered to evaluate the overall human-intelligent system interaction:

1. Is the necessary information for the human to be able to determine that an intervention is needed present?
2. Is the information presented in an appropriate form?
3. Is the interaction language efficient for both the human and the intelligent system?
4. Are interactions handled efficiently and effectively - both from the user and system perspective?
5. Does the interaction architecture scale to multiple platforms and interactions?
6. Does the interaction architecture support evolution of platforms?

The first four issues are relevant to all intelligent systems. If we are concerned with supporting 1:N and N:1 interactions, we must evaluate the scalability of the interaction architecture. If we are interested in using the architecture over a period of time, we must consider how the evolving behaviors of new intelligent systems will be supported.

Usability evaluations of desktop software products use three metrics: effectiveness, efficiency, and user satisfaction [10]. Due to the dynamic nature of intelligent systems separating the evaluation into two pieces, getting the proper information to the user and the actual performance of the user/system in the interaction, produces a finer granularity of understanding. Users may have all the information they need but the interaction can fail for other reasons. Likewise, the interaction may be successful without users having the proper information. By separating the evaluation into these pieces, we reduce the risk of counting these cases in the results.

These evaluation questions cannot presently be answered in a general sense. Our approach is to narrow both the domain and the role of the human interaction and systematically explore the space. After we have explored a number of roles and domains of interaction, we will examine the results to determine if there are commonalties that can be expressed as guidelines for interaction guidelines.

3. Evaluation Methodologies

The six issues listed above are evaluated using different types of evaluations. In this section we discuss the types of evaluations appropriate for each issue.

3.1. Information Presence and Presentation

To determine if the necessary information is presented – and in the correct form – we are customizing a situational

awareness assessment methodology. Situational awareness [6] is the knowledge of what is going on around you. The implication in this definition is that you understand what information is important to attend to in order to acquire situational awareness. Consider your drive home in the evening. As your drive down the freeway and urban streets there is much information you could attend to. You most likely do not notice if someone has painted their house a new color but you definitely notice if a car parked in front of that house starts to pull out in your path.

Level One of situational awareness (SA) is the basic perception of information in your surroundings. For example, in driving did you notice the cars to the left, right, front and rear of your vehicle? Failures to perceive information can result as short comings of a system or they can be due to a user's cognitive failures. In studies of situational awareness in pilots, 76% of SA [11] errors were traced to problems in perception of needed information. Level Two of situation awareness is the ability to comprehend or to integrate multiple pieces of information and determine the relevance to the goals the user wants to achieve. A person achieves the third level of situational awareness if she is able to forecast future situation events and dynamics based on her perception and comprehension of the present situation.

The most common way to measure situational awareness is by direct experimentation using queries [5]. The task is frozen, questions are asked to determine the user's situational assessment at the time, then the task is resumed. The Situation Awareness Global Assessment Technique (SAGAT) tool was developed as a measurement instrument for this methodology 7]. The SAGAT tool uses a goal-directed task analysis to construct a list of the situational awareness requirements for an entire domain or for particular goals and sub-goals. Then it is necessary to construct the query in such a way that the operator's response is minimized. For example, if a user were being queried about the status of a particular robot, the query might present the robot by location rather than replying on the user to recall a name or to understand a description. The various options for status could be presented as choices rather than relying on the user to formulate a response that might not include all the variables desired. SAGAT queries are constructed to include measures of all three levels of situation awareness. Queries related to placement of nearby vehicles in a driving domain would measure level one situation awareness. Queries that ask about vehicle activity, such as cars that have just switched lanes or increased or decreased speed would address level two. Level three, prediction, would be measured by queries such as the likelihood that the car in front of you will move into the far left lane. This could be determined by the observation of a turn signal or the previous pattern of behaviors of lane switching by that automobile.

3.2. Interaction Performance

Information presence and presentation evaluations are user-centric. Interaction performance evaluations need to take into account the performance of both the human and the intelligent system. We are concerned with measuring the ability of the user to formulate the correct interaction and the system to understand and carry this out. This type of evaluation can be conducted as a typical HCI evaluation [4]. To elaborate, a set of tasks are constructed and explained to the user. The user is then directed to use the interface to accomplish each task. The training and the expertise of the user can confound this evaluation. In general, the users should be chosen from the representative population of users and given the same amount of training as those users would be given. If there are a number of diverse users, then different classes of users should be identified and between five and eight users from each class should be used in the evaluation. From the user perspective the metrics should reflect the number of tasks that the user is able to successfully complete, the time for each task, and a user satisfaction measure that can be obtained using a standard questionnaire [16].

The system performance is also factored into this metric. The effectiveness measure has two components – the user executing the correct interaction and the system responding correctly to this interaction. Likewise the efficiency metric would be the sum of the user time for the interaction to be specified and the system time for it to be carried out.

Currently we have not considered the notion of mixed initiative interaction; that is, the intelligent system notices that an interaction by the user is needed and notifies her of this. This would require ensuring that these notifications are seen and understood by the user (the situation awareness measurement) and that the correct interaction is selected and carried out (the interaction performance portion).

3.3. Support for Scalability and Evolution

Support for 1:N and N:1 interactions should be evaluated using the information presence and presentation, and performance methods. The information to be displayed in each case and the presentation of that information needs to be determined and evaluated using a situational awareness assessment. The performance measures need to ensure that the user can identify the appropriate interaction for the appropriate platform within the appropriate time. This will be critical if a number of heterogeneous platforms are being used. In the case of

multiple people interacting with the system simultaneously, it will be interesting to determine how to display that information and what effect this will have on the interaction of any given role. For example, how will the operator behavior change if a number of bystanders are present when she needs to teleoperate a robot to get it into a building?

Evaluating the interaction support for evolution is more difficult. Intelligent systems will evolve and be capable of undertaking more tasks successfully and communicating with users at higher levels of abstraction. This will definitely necessitate re-examining the interaction language, but the information that is needed has to be reconsidered as well. A more robust level of autonomy might be supported as effectively using more abstractions in the first level of information presented.

3.4 Evaluation Methodologies for Different Roles

The five roles defined here, supervisor, operator, mechanic, teammate, and bystander, clearly have different information requirements and different interactions. Is it feasible to use the same type of evaluation to measure the performance of all roles?

We expect that the supervisor, operator, and mechanic will have access to a specialized display of information from the intelligent system. This display may be on a workstation, laptop or a small handheld device. However, this display will give the appropriate situational awareness to the users. Teammates may not have access to such a device and bystanders certainly will not. A different methodology is needed in these situations. This will be discussed in section 5.

All of the interaction roles will have access to some subset of the interaction vocabulary. A particular interaction may be selected using a typical command language, keyboard input, voice, or pen-based types of interactions. The action might even be initiated by using "physical" manipulation. For example, moving in front of a robot and touching a sensor on the robot to cause an action to occur would be an example of "physical" manipulation. The same performance evaluations will be used in all roles. However, in cases where no specialized display is available, the challenge will be to make the interaction choices known to the user.

4. Case Study One: Developing the Situational Awareness Assessment Tool

In our current work, we have narrowed the domain to a mobile, robotic platform that is given the task of driving from one place to another in a complex urban environment. The human-robot interaction role is that of the supervisor, similar to a driving coach with a student behind the wheel. As the capabilities of the student driver increase, the driving coach should be able to pay less attention and only give guidance when she detects the possibility of a problem. However, our driving coach will be remote. For the first part of the work we are investigating the first two questions: what information is necessary for the human to decide that intervention is needed and what is the appropriate presentation of this information?

The first step is to determine what information is needed by the user. In the driving domain this task is easier than in most because of the amount of information available. There are numerous studies on driving [15] and our own experience that we will turn to for the initial design of the user interface.

We currently characterize the information using four categories: static environment; dynamic environment; platform information, and task information. The static environment consists of information in the environment that does not change or at least changes very infrequently. In the driving domain, this would be the location of cross streets and intersections; the type of roadway; whether there are stop signs, stop lights, or other traffic controls. Dynamic environment information examples are the amount of traffic present, the pedestrians traffic if the driving environment is urban at the time, and the status of the traffic light. Examples of vehicle information are the speed of the vehicle, the amount of fuel left, and the condition of the vehicle such as non-working turn indicators. Task information is the knowledge of the destination; the current distance to the destination; how far to the next decision point.

For given situations, the information needs change. For example, approaching a green light, drivers should look for different hazards than when approaching a red light.

Once the information for selected situations is determined and the user interface is designed, the awareness assessment tool is constructed. As explained in the earlier section, this is accomplished by using a simulation and freezing the simulation at a certain point. The simulation screen is blanked out and the user is directed to answer a series of questions to determine what her situational awareness is at this time. The queries that are constructed are the most important aspect of the assessment methodology. Again, expert elicitation in some form is used to obtain this information. This can be done by observations of performance, verbal protocols, interviews, or questionnaires. The results can be combined and later verified by a number of subject experts. In the driving domain, we are utilizing a tutorial used to teach driving [16]. The tutorial presents a number of situations and the student is asked to identify potential

hazards or take action to prevent accidents. The information given is the front view, the rear and side view mirrors and the instrument panel. The accompanying instructor manual identifies the information that students needed to have identified. We will use this information to construct our queries.

The presentation of the queries should be done so that the user can quickly answer. The users should not be asked to recall information that is not relevant to the situation. For example, asking a user to designate if there are cars to the left, right, front (within 2 car lengths) and back (within 2 car lengths) of his car on a multilane highway is fine. Requesting the number of cars for most other situations is not fine.

The analysis will be done for each situation that we present and for the different information classifications that we have identified. We will use situations from highway driving, urban driving, and illustrating normal conditions as well as hazardous conditions.

We intend to use our user interface and the situational awareness assessment results as a baseline. These will be made public. Others interested in this particular domain could either construct a new user interface to display the same information we have identified and compare their results. Alternatively, different information could be displayed in the user interface and the results compared with a baseline.

Once we have completed our information presence and presentation evaluation we will proceed to the performance evaluation. We intend to also look at scalability of the user interface from one to multiple robotic platforms.

5. Case Study Two: Examining User Mental Models in the Bystander Role

We are also working at the opposite end of the spectrum and looking at evaluation methodologies where no specialized visual user interface is present. We have designed an experiment to examine the effects of consistency and expectedness of behavior on bystanders' abilities to construct a mental model of the robot's capabilities. We are using a Sony Aibo ™ [1] for this experiment and as the robot has a dog-like appearance we have designed behaviors that one would normally expect of a dog (playing with a ball, sitting) and others (singing, dancing) that would not be expected.

[1] * The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

Observations of the users as they interact with the robot will be recorded and users will be asked after the experiment to identify what robot behaviors will result from various interactions on their part. Results from the first round of this study will be available this fall.

6. Conclusions

We have defined five different interaction roles for humans and intelligent robotic systems. We have also defined six issues that constitute "performance of intelligent systems." We have outlined the evaluation methodologies that can assess these measures of performance. We are currently conducting evaluation experiments for two types of interaction roles in two domains: the supervisory role in a driving domain and a bystander role in a social interaction domain. Our future plans are to use the framework suggested by roles and domains to systematically explore evaluation methodologies for human-robot performance. Our tools and results will be made publicly available as our research progresses.

7. Acknowledgements

This work is funded by DARPA, contract number E710. The author would like to thank Dr. James Albus and Elena Messina for their support and encouragement in including the human element in their research programs. Brian Antonishek, NIST and Siavosh Bahrami, University of California, Irvine are working on the experiments discussed in this paper.

References

1. AAA Foundation for Traffic Safety. 2002. driver-ZED.
2. Albus, J. 1997. 4-D/RCS: A Reference model Architecture for Demo III, NISTIR 5994, Gaithersburg, MD. March.
3. Brown, M. and Leveson, N. 1998. Modeling Controller Tasks for Safety Analysis. Presented at the Workshop on Human Error and System Development, Seattle, April.
4. Dumas, J. S. & Redish, J. C. 1993. A practical guide to usability testing. Norwood, NJ: Ablex Publishing.
5. Endsley, M. R., 2000. Direct Measurement of Situation Awareness: Validity and Use of SAGAT in (Eds) Mica R. Endsley and Daniel J. Garland . Situation Awareness Analysis and Measurement, Lawrence Erlbaum Associates: Mahwah, New Jersey.
6. Endsley, M. R., 2000. Theoretical Underpinnings of Situation Awareness: A Critical Review, in (Eds) Mica R. Endsley and Daniel J. Garland . Situation Awareness Analysis and Measurement. Lawrence Erlbaum Associates, Publishers: Mahwah, NJ
7. Endsley, M. R. 1988. Design and evaluation for situation awareness enhancement. In Proceedings of the Human Factors Society 32nd Annual meeting (Vol.

1, pp 97-1010 Santa Monica, CA: Human Factors Society.
8. Fong, T., Thorpe, C. and Bauer, C. 2001. Collaboration, Dialogue, and Human-robot Interaction, 10th International Symposium of Robotics Research, November, Lorne, Victoria, Australia.
9. Gugerty, L. 1997. Situation Awareness during Driving: Explicit and Implicit Knowledge in Dynamic Spatial Memory. Journal of Experimental Psychology: Applied. Vol. 3 (1), 42-66.
10. ISO 9241-11. 1998. Ergonomic requirements for office work with visual display terminals (VDT)s part 11 Guidance on usability.
11. Jones, D.G. and Endsley, M. R. 1996. Sources of situation awareness errors in aviation. Aviation, Space and Environmental Medicine 67(6), pp. 507-512.
12. Leveson, N. de Villepin, M., Daouk, M. , Bellingham, J., Srinivasan, J., Neogi, N., Bachelder, E., Pilon, N., and Flynn, G. 2001. A Safety and Human-Centered Approach to Developing New Air Traffic Management Tools. ATM 2001, Albuquerque NM, December
13. Leveson, N. and Palmer, E. 1997. Designing Automation to Reduce Operator Errors In the Proceedings of Systems, Man, and Cybernetics Conference, Oct.
14. Norman, D. 1986. Cognitive Engineering in Donald Norman and Stephen Draper (Eds.) User-centered design:new perspectives on human-computer interaction, Erlbaum Associates: Hillsdale, N.J, 31-62.
15. Picard, R. (1997), Affective Computing, MIT Press, Cambridge.
16. The Software Usability Measurement Inventory. (SUMI) Accessed July 12, 2002. http://www.ucc.ie/hfrg/questionnaires/sumi/
17. Scholtz, J. (2002) Creating Synergistic CyberForces in Alan C. Schultz and Lynne E. Parker (Eds.), Multi-Robot Systems: From Swarms to Intelligent Automata. Kluwer.