Human-Robot Interaction: Development of an Evaluation Methodology for the Bystander Role of Interaction[*]

Jean Scholtz
National Institute of Standards and Technology
MS 8940
Gaithersburg, MD 20899
Jean.scholtz@nist.gov

Siavosh Bahrami
4 Falling Leaf
Irvine, Ca 92612
siavoshb@yahoo.com

Abstract - *Various methods can be used for evaluating human-robot interaction. The appropriateness of those evaluation methodologies depends on the roles that people assume in interacting with robots. In this paper we focus on developing an evaluation strategy for the bystander role. In this role, the person has no training in interacting with the robot and must develop a mental model to co-exist in the same environment with the robot.*

Keywords: Human- robot interaction, social interaction, user roles, mental models, conceptual models, intelligent systems.

1 Introduction

Robots are moving out of the research laboratory and into society. They have been used by the military to search caves in Afghanistan [5]. They were used in search and rescue at the World Trade Center [8, 11, 19]. Robots have been introduced as toys [2, 17] and household tools [14]. Robots are also being considered for use in domains such as elder care [1,12]. As robots become more a part of our society, the field of human–robot interaction (HRI) becomes increasingly important. To date, most interactions with robots have been by researchers in robotics, in their laboratories. Now we expect people with real-world tasks to interact with these robots for work and play. How do we design and evaluate the user interfaces and interaction techniques for human-robot interaction?

What is a robot? A web search for a definition of a robot reveals several types: knowledge robots (commonly referred to as "bots"), computer software robots that continuously run and respond automatically to a user's activity, and industrial robots. A dictionary definition [Collins English dictionary] of the noun 'robot' is "any automated machine programmed to perform specific mechanical functions in the manner of a man." Murphy [7] defines an intelligent robot as a mechanical creature that can function autonomously. She notes that while a computer may be a building block of the robot, the robot differs from a computer in that it can interact in the physical world by moving around and by changing aspects of the physical world.

It follows that human-robot interaction is fundamentally different from typical human-computer interaction (HCI). Fong et al. [6] note that HRI differs from HCI and Human-machine Interaction (HMI) because it concerns systems that have complex, dynamic control systems, exhibit autonomy and cognition, and 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 interact with simultaneously; the degree of autonomy of the robot; and the environment in which the interactions occur.

2 Roles of Interaction

Scholtz [15] defines three different roles for users interacting with robots: supervisor, operator, and peer. A subsequent paper [16] expands these roles into five distinct interaction categories. The operator role has been subdivided into an operator and a mechanic role. The peer role has also been subdivided into a bystander role and a teammate role. Supervisors are responsible for overseeing a number of robots and responding when intervention is needed – either by assigning an operator to diagnose and correct the problem or assisting the robot directly. The operator is responsible for working "inside" the robot. This might involve assigning way points, tele-operating the robot if needed, or even re-programming on the fly to compensate for an unanticipated situation. The mechanic deals with hardware and sensor problems but must be able to interact with the robot to determine if the adjustments made are sufficient. The teammate role assumes that humans and robots will work together to carryout some task, collaborating to adjust to dynamic conditions. The bystander would have no formal training with the robots but must co-exist in the same environment with the robots for a period of time and therefore needs to form some model of the robot's behavior. Some of these roles can be carried out remotely as well as locally. In order to evaluate HRI we need to consider the role or roles that individuals will assume when interacting with a robot.

For example, our hypothesis is that supervisors need situational awareness of the area and need to monitor both dynamic conditions and task progress. An operator, on the other hand, needs to have knowledge of the current mode of the robot, the condition of any sensors, and an awareness of any obstacles in close proximity to the robot. The mechanic would be aided by having access to logs of behaviors to troubleshoot the problem. Users may or may not have a remote interface for a robot teammate. They will certainly use gestures and verbal commands to interact [13] but they need some confirmation that the robot has understood the command and is able to carry it out. Bystanders will not have any experience with a particular robot and will need enough information about what the robot can do and is doing to feel comfortable in the shared environment. In addition, if multiple people are interacting in different roles with the same robot, some level of awareness of these interactions may be necessary.

3 Evaluation of Human-Robot Interaction

Typical HCI evaluations use efficiency, effectiveness, and user satisfaction as measures when evaluating user interfaces. Effectiveness is a measure of the amount of a task that a user can perform via the interface. Efficiency is a measure of the time that it takes a user to complete a task. Satisfaction ratings are used to assess how the user feels about using the interface. These three measures seem appropriate for evaluation of a number of HRI roles. The roles of supervisor, operator, mechanic, and team mate will all involve some sort of task and can benefit from using efficiency, effectiveness, and satisfaction as metrics. Additionally, because robots interact with the physical world and may at times be remote from the user, the user will need some awareness of the robot's current situation. This involves both an understanding of the external environment as well as the internal status of the robot. Additionally, some roles such as the team mate assume that the user is performing other tasks as well as interacting with the robot. Workload measures, such as the NASA Task Load Index (TLX) [9] , can be used to determine the load that the HRI places on the supervisor or operator of the robot.

The bystander role, however, will not involve performing specific tasks with the robot. Rather we envision the bystander role as an understanding of what the robot can do in order to co-exist in the same environment. Consider the following examples.

3.1 Robots as pets in an elder care facility

You are going to visit your aunt for the afternoon. You find her playing with her robot dog. Your aunt has some memory problems and she is having difficulty remembering how to get the dog to do some of its tricks. She asks you to help. How do you determine what the dog can do? Most likely you use trial and error. But what affects your chances of success in building up a model of what the robot can do?

3.2 Driving on the same road as an autonomous vehicle

You are driving along the freeway and you notice that no one is seated behind the wheel of the vehicle next to you. After a short time you notice that the traffic ahead of you is slowing down and you see that road work is blocking your lane. Cars ahead of you are merging into one lane. You should be able to merge in front of the autonomous vehicle. How comfortable do you feel doing this?

4 SOCIAL INTERACTION

The bystander roles falls into an existing category of research described as social interaction. Research in this area has concentrated on understanding social gestures and vocalizations that humans use in their communications with each other and modeling this behavior in software for robotic systems. Brezeal [3] looks at language interaction but focuses on tones of the voice rather than content of the language interaction. The robot senses the user's tone of voice and matches it's facial expressions and speech tone to that of it's user.

Nass et al [10] explored the effects of various embodiments for conversational agents. This research looked at the ethnicity and personality of conversational agents and assessed user satisfaction in interacting with agents belonging to the same or different group as the participants. When participants and the conversational agents were of the same ethnicity, the participants found the agents more socially attractive and trustworthy. To investigate personality affects, agents were designed to be introverted or extroverted. Personality cues given by the agents were both verbal and nonverbal. The experiment manipulated the consistency of the verbal and nonverbal cues with the personality type of the agent. Participants liked the consistent behavior of the agent and found it more fun to interact with. However, they liked the character whose nonverbal cues more closely matched their own personality type.

Research on interactive toys may also be helpful in developing HRI evaluations for the bystander role.

Strommen [17] performed a number of studies to design ActiMates Barney, an animated plush doll that could be used either as a free-standing toy, in conjunction with a TV or video player, or connected to a computer. Based on his research Strommen noted some guidelines for the design of interactive toys.

1. The toys should be friendly but should give the children directives as opposed to using questions to interact.
2. Each sensor on the toy should be associated with one function. Children were not able to use combinations of sensors to produce actions. However, the different sensors did have a series of actions that were produced. For example, pressing the feet of A/Barney caused a song to be sung. But which song was sung at any time was random. Children did try to press the feet a number of times to bring up a particular song.
3. Children also want to be able to interrupt the action by interacting with a different sensor. The model used originally in the design was that children would play along with the animated toy. However, children clearly showed that they wanted to be in control and have the toy play along with them.
4. Because A/Barney had three different modes of interaction (standalone, with TV, with computer) making the functions consistent across all modes was an issue. This was accomplished by using the same basic functionality but making the functionality appropriate to the social context of the situation.

5 DEVELOPING AN EVALUATION METHODOLOGY FOR THE BYSTANDER ROLE IN HRI

Implicit in the research of both Stommen and Nass is that users were building a mental model or conceptual model of what the interactive object did. Mental models [4,18] or conceptual models provide the basis for understanding an interactive device or program. It names and describes the various components and explains what they do and how they work together to accomplish tasks. Understanding the conceptual model makes it possible to anticipate the behavior of the application, to infer “correct” ways of doing things, and to diagnose problems when something goes wrong. . Users of computing systems build “appropriate” mental models [18]. That is, models that are useful in explaining behaviors. Note that these mental models are not complete models and in many instances may even be erroneous.

Designers have a conceptual model that they use in producing the device. Users build up a conceptual model as they interact with the device. Desktop computing applications should be designed to support the acquisition of appropriate conceptual models. Analogies or metaphors, such as the desktop metaphor, facilitate the user in building conceptual models. A robot with no visual display and whose behaviors may change depending on the context of the environment make it challenging for users to build unified models of behaviors and interactions. We proposed an experiment to assess the conceptual model of HRI that users were able to build after a short interaction period with the robot. We used the following four metrics in our initial experiment:

1. Predictability of behavior

   Metric: degree of match between user’s model of behavior and actual behavior of the robot.
   For example, how many behaviors performed by the robot is the user able to predict ? Given a particular interaction with the robot is the user able to predict the response ?

2. Capability awareness

   Metric: degree of match between user’s model and the actual functionality of the robot.
   Does the user have a model of all the possible behaviors that the robot is capable of ?

3. Interaction awareness

   Metric: degree of match between user’s model and the actual set of interactions possible.
   Does the user understand all the ways to interact with the robot ?

4. User satisfaction

   Metric: rating scale or responses to questions about interactions.
   How satisfied is the user with the interactions ?

6 EXPERIMENT

We designed the experiment to have two stages. In the first stage we investigated interaction awareness. In the second stage we assessed predictability of awareness and capability awareness. In our post-experiment debrief we looked at user satisfaction.

For this initial experiment we used the Sony AIBO ™[1]. Figure 1 shows the robot that we used in the experiments.

Figure 1 : Sony's AIBO 220E was used in the study

In order to test the sensitivity of our metrics, we manipulated the behavior of the robot. The AIBO has dog-like appearance and we hypothesized that its form would be a factor in the bystander's expected capabilities. We implemented two sets of behaviors, one consistent with a dog-like behavior, and another with non dog-like behaviors. We subdivided each set of behaviors into consistent and inconsistent behaviors. The consistent set would produce the same action each time the user performed the matched interaction. The inconsistent behavior produced one of a set of behaviors selected randomly from 4-6 different behaviors. Figure 2 gives some examples of the dog-like and non dog-like behaviors.

Behavior type	**Examples**
expected, consistent (EC)	walking; playing with a pink ball; sitting down
unexpected, consistent (UC)	talking; dancing; waving
expected, inconsistent (EI)	same as expected, consistent but with a certain degree of random behavior
unexpected, inconsistent (UI)	same as unexpected, consistent but with a certain degree of random behavior

Figure 2: Examples of the behavior sets used in the experiment

There are three ways to interact with the Sony AIBO. Speaker independent voice recognition can be used to give voice commands. The dog has buttons on its back and head that can trigger behaviors. A camera in the dog's head can be used to trigger behaviors based on visual interaction. We used all of the methods in our study. We used 5 voice commands, 5 buttons, and a visual interaction in which the robot responded if it was shown a pink ball.

7 ACTUAL EXPERIMENT

We had 20 participants in our study. They were randomly assigned to one of the four behaviors, giving us five subjects for each behavior set. The participants in the study were all between the ages of 19 and 25, evenly split between males and females. All were undergraduates participating in a summer research program at the National Institute of Standards and Technology. Because we were testing the methodology and not focused on results we were more concerned with having a homogeneous set of participants. When we actually conduct the experiments we will need to select a larger and more heterogeneous group.

Participants were first asked a few demographic questions. The students were all working in some area of science. Six of the participants were involved in some aspect of computer science with the other fourteen studying physics, chemistry, mechanical engineering, etc. Five of the participants had some experience with robots – mostly as interactive toys that were very limited in what they could do.

Participants were asked how they thought they could interact with the robot and we recorded their answers to determine interaction awareness. We then told them the interactions that they could use. We asked the participants to play with the robot for 10 minutes to get an idea of what it could do and we observed their interactions. After the time was over we asked the participants to tell us what the robot did in response to the different interactions. We recorded this information to measure predictability of behavior and capability awareness.

Table 1 shows the results of our initial assessment of interaction awareness. Participants could see the robot at this point in time but were asked not to try interacting with it yet. In addition to the three types of interactions possible with the robot, participants also thought it might be possible to interact by touch – specifically petting, by using some sort of remote control or infrared device, or possibly the robot might use smell to identify people and objects. Table 2 shows the number of participants who correctly identified one or more interaction modes for the AIBO.

Interactions	**Number of participants**
Voice	11
Buttons	6
Vision	10
Touch/ pet	4
Remote control	7
Smell	1

Table 1: Types of interactions that participants expected

[1] The identification of any commercial product or trade name does not imply endorsement or recommendation.

Number of interactions correctly identified	Number of participants
0	3
1	9
2	6
3	2

Table 2: Number of participants who correctly predicted interaction modalities.

Figure 3 shows the results of the predictability of behavior indicator. There were 11 interactions provided. We asked participants after their 10 minutes of playing with the robot to tell us what behaviors resulted from each interaction. For the consistent behavior we scored the response as a 0 if the participant gave no answer, 1 if the answer was partially correct, and 2 if the answer was completely correct. For the inconsistent behaviors we scored 0 for no answer, 1 if the participant mentioned 1 or more behaviors, and 2 if the participant mentioned some degree of randomness in the behaviors. Each bar in Figure 3 corresponds to one participant's score. The maximum score that could be obtained was 22.

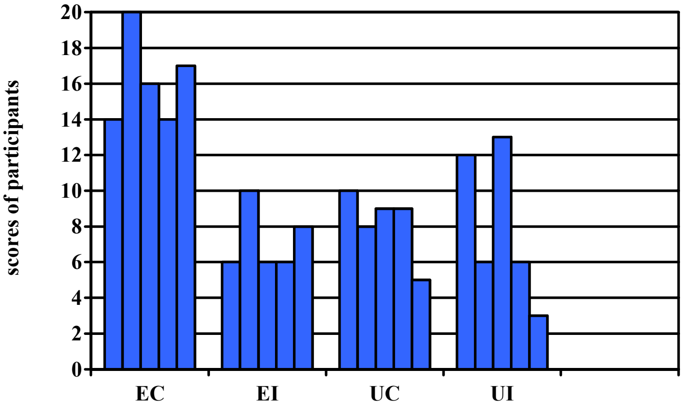

Figure 3: Accuracy of conceptual models of participants for each behavior set.

We also asked participants if they enjoyed interacting with the robot and asked them if their expectations had changed based on their interactions. Sixteen of the participants said they enjoyed the interaction. Two participants said they enjoyed interacting for a short period of time. Two other participants said it was boring or frustrating.

Positive comments from participants mentioned the use of voice interaction. Several were impressed with what the toy could do. Participants used adjectives like cool, amazing, high tech to describe the robot. Negative comments expressed disappointment with what the robot could do, wanted more dog-like behaviors, and better voice understanding. Several participants also wanted the robot to accept multiple commands at a time and the ability to cancel a command.

8 DISCSSION OF RESULTS

While our focus in this experiment was on developing the methodology for evaluating the bystander role in HRI, some of our observations of interactions may be useful in refining the methodology or for suggesting additional metrics.

Testing interactions poses a problem when the interaction technology is not as robust as it should be. In our experiment, voice recognition was a problem. One participant in particular had a distinct accent and was unable to get the voice commands to work. In general, participants tolerated some errors on the part of the voice recognition saying it was just like their dog at home. However, errors in interaction modalities will certainly hinder participants in creating conceptual modes.

In both sets of unexpected behaviors (UC and UI), participants asked how they could get the robot to do dog-like things. They were frustrated because the dog didn't walk or follow the pink ball. Several participants tried to say dog-like commands to the robot, such as "sit" or "fetch". In addition to asking participants what they think capabilities are, recording these interactions and noting the percentage that are "out of scope" for the robot can be used to measure capability awareness.

In general, participants who received the unexpected behavior treatments seemed more frustrated. Also, participants in the inconsistent behavior sets were reluctant to say that the behaviors were random or inconsistent. A number of the participants blamed themselves, saying that they weren't very good at figuring this out. We certainly will use a frustration rating in our user satisfaction scale. These observations also suggest that the predictability of behavior metric might be accompanied with a confidence level.

Participants in general had difficulties figuring out when a behavior had ended. In particular, the robot was programmed to find and move to the pink ball when it was visible. Some participants had difficulty in determining that the behavior ended only when they moved the pink ball out of sight. Participants also tried to overlap behaviors. They tried to give the robot verbal commands while it was still executing another behavior. This is similar to the desired for interruption that Stemmen found in his studies. This suggests that some rating of the amount of user control is desirable. Also, we intend to factor such attempts into our measure of capability awareness.

9 CONCLUSIONS

We are interested in continuing our research in this area and intend to use our results from this exploratory study to refine our methodology as well as our hypotheses. Refinement is needed in several areas.

First, interaction awareness needs to be measured at a finer level. We were able to determine the interaction modes that participants were aware of, but we didn't assess what voice interactions participants believed they could issue. We need to separate out capability awareness from predictive behavior. For the next experiment, we will ask participants what type of actions they think the robot can do before the interaction period.

It was difficult to make sure that participants tested all the interactions. As two sets of behaviors contained random interactions we need an accurate way of logging what interaction-action pairs participants saw. We intend to implement a logging capability on the robot to record this information. Based on our observations during this pilot study we intend to develop ratings for user satisfaction to use along with participants' responses to more open ended questions.

As we did see differences in the accuracy of the conceptual models between the different sets of behaviors, we believe that the methodology for measuring predictive behavior is appropriate.

10 ACKNOWLEDGMENTS

This work was supported in part by the DARPA MARS program. Siavosh Bahrami held an NSF/NIST Summer Undergraduate Research Fellowship during the time this work was accomplished.

11 REFERENCES

[1] ABC News.com, http://more.abcnews.go.com/sections/wnt/dailynews/robots_elderly020409.html, accessed August 28th, 2002.

[2] AIBO, http://www.aibo.com/, accessed August 28th, 2002.

[3] Breazeal, C. 2000. "Sociable Machines: Expressive Social Exchange Between Humans and Robots". Sc.D. dissertation, Department of Electrical Engineering and Computer Science, MIT.

[4] Carroll, J. And Olson, J. 1988. Mental Models in Human-Computer Interaction. In M. Helander (ed.) *Handbook of Human-Computer Interaction.* Amsterdam : Elsevier Science Publishers B.V. (North-Holland). 45-61.

[5] Christian Science Monitor, http://www.csmonitor.com/2002/0731/p01s03-usmi.html, accessed August 20, 2002

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

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

[8] Murphy, R., Blitch, W., and Casper, J. 2002. AAAI/Robocup-2001 Urban Search and Rescue Events: REality and Competition, AI Magazine, 23(1), Spring 2002.

[9] NASA Task Load Indecx (TLX). http://iac.dtic.mil/hsiac/Products.htm. accessed May 29, 2003.

[10]Nass, C, Isbister, K. and Lee, E. 2000. Truth is Beauty: Researching Embodied Conversational Agents in J. Cassell, J. Sullivan, S. Prevost, &E. Churchill (eds), Embodied Conversational Agents.

[11]National Geographic News, http://news.nationalgeographic.com/news/2001/09/0914_TVdisasterrobot.html, accessed August 20, 2002.

[12]Nursebot Project, http://www-2.cs.cmu.edu/~nursebot/, accessed August 28th, 2002.

[13]Perzanowski, D., Schultz, A., Adams, W., Marsh, E., and Bugajska, M. 2001. "Building a Multimodal Human-Robot Interface," Intelligent Systems, 16(1), Jan/Feb 2001, IEEE Computer Society, 16-21.

[14]Robotic Mower, http://www.shopping-emporium-uk.com/mower/, accessed August 28th, 2002

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

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

[17]Strommen, E. 1998. When the Interface is a Talking Dinosaur : Learning Across Media with ActiMates Barney. In *Human Facotrs in Computer Systems. Proceedings of the ACM SIGHCHI Conference* (Los Angeles, April 1998), ACM Press, 288-295.

[18]Van der Veer, G. And Melguizo, M. 2003. Mental Models. In J. Jacko and A. Sears (Eds). The Human-Computer Interaction Handbook. Mahway, New Jersey : Lawrence Erlbaum. 52-80.

[19] Wired Archive, http://www.wired.com/wired/archive/10.05/robots.html, accessed August 20, 2002.