Machine Trait Scales for Evaluating Mechanistic Mental Models of Robots and Computer-Based Machines

Sara Kiesler and Jennifer Goetz, HCII,CMU

April 18, 2002

In previous work, we and others have used the Big Five Personality Scales and other human trait rating scales to evaluate anthropomorphism and social responses to machines (see our CHI short papers). However, we believe that most people correctly perceive that machines are inanimate and not "real" people. In other words, anthropomorphism is partial. Current research in cognitive science implies that when a machine, such as a robot, exhibits humanlike behavior (such as speech and ostensibly intentional movement), people retrieve cognitions associated with people and with machines from long-term memory. From these cognitions, people create a more or less coherent mental model of the machine they are observing.

The purpose of this research was to create rating scales for robots (and other computer-based machines) that could be used to measure the extent to which people's mental model of the technology incorporates inanimate and mechanistic elements. We plan to use these scales in conjunction with measures used in studies of person perception, such as the Big Five personality scales and intellectual evaluation items. Whereas the latter measure animist and anthropomorphistic elements of a person's mental model of a machine, our new scales will measure the inanimate and mechanistic elements of the mental model.

Method

We first created a pool of 63 traits from several sources: Handbook of Human Factors Design (Woodson et al, 1992), a study of seven people who were asked to photograph and describe their favorite appliances, tools, and machines at home and at work, and through brainstorming in our research group. We then combined these traits with trait names from the Big Five Inventory (John et al, 1991) and intelligence evaluation items (Warner & Sugarman, 1986) to create two versions of a questionnaire. One version asked participants to rate the traits on 5-point scales in answer to the statement: "Please rate each of the attributes below and say how human-like they are." The other version asked participants to respond to the same rating scales in answer to the statement: "Please rate each of the attributes below and say how machine-like they are." We then compared the responses of 20 students at Carnegie Mellon University who answered the first version with the responses of 20 students who answered the second version. For each trait, the mean difference between the responses is assumed to reflect the degree to which people perceive each trait to be characteristic of humans or of machines. For instance shyness is considered a highly humanlike trait, while breakable is considered machinelike. Traits such as dependable and dangerous, however, are applicable to both humans and machines. Table 1 lists all traits receiving an average machinelike score significantly greater than its humanlike score.

We next chose 33 traits from the longer list, all of which had received ratings significantly greater on machinelikeness (see Table 1). Sixty participants were recruited for an online Web survey in which we asked them to rate the Nursebot project's Pearl robot and one of the following machines: a car, a personal computer, and the Pyxis robot. Because our project involves a comparison of humanlike robots with other types of robots, we wanted every participant to rate a humanlike robot, to provide a more humanlike machine as an anchor in doing the other machine ratings (see Figure 1).

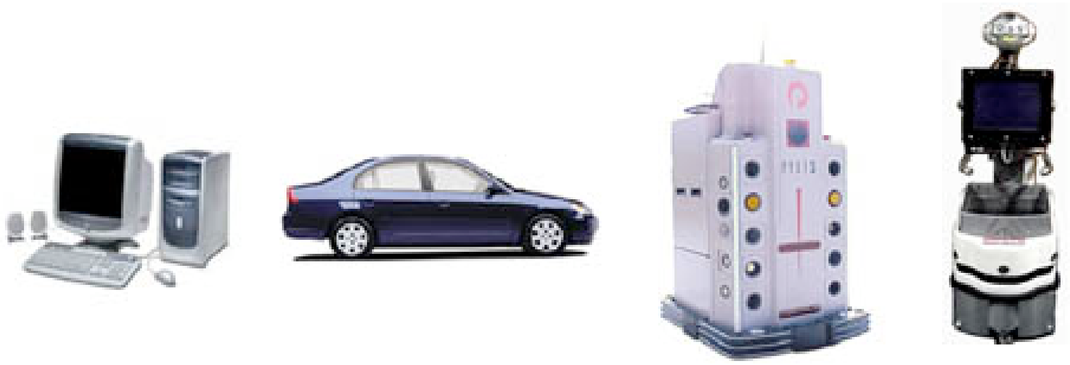

Figure 1. Objects rated in the online Web study. From left to right: a personal computer, a car, the Pyxis robot, and Pearl.

Results

In Table 2 we present the results of the factor analysis using ratings of the Pearl and Pyxis robots and a personal computer.[1] The principal components analysis performed as a first step resulted in 8 components with Eigenvalues greater than 1.0 and accounting for 69% of the variance, but we set the number of factors to equal 5 (57% of the variance) because more factors did not result in easily interpretable scales with multiple items.

Creation of scales

We selected items from each factor that loaded .5 or more on the factor to create each scale. We eliminated a few items because they reduced scale reliability. The resulting scales for measuring people's perceptions of robots and computers are shown in Table 2. The five scales measure dimensions we call Efficiency, Maintenance, Durability, Safety, and Information Technology.

Cross product comparisons

The mean scale ratings of the PC, Pearl robot, and Pyxis robot are shown in Figure 2. A repeated measures analysis of variance indicated that, overall, the three machines were rated differently (F [2, 85] = 27, $p < .001$) and there was an interaction of scale with machine rated (F [8, 164]= 14.8, $p < .001$). Generally, the PC was rated especially highly on the Efficiency Scale ($p < 01$). The Pearl robot was viewed as high maintenance ($p < .01$). The Pyxis robot was rated very high on the Durability Scale, with the PC next and the Pearl robot much lower ($p < .001$). The PC was perceived as safest ($p < .001$). Finally, there were not differences among the PC and the two robots on the information technology scale. All of these findings appear to support the face validity of the scales.

[1] Analyses using the ratings of the car produced similar but results, but items related to safety, hazards, and durability loaded much higher on the scales.

Anchoring effects

Research in social psychology has shown that people's ratings of any object or person are affected by the context in which they are rated, especially by the implicit comparison objects or persons. For example, a tall person will be rated taller when imagined next to a short person. In our earlier work, we have suspected that participants' personality and intelligence ratings of a robot were very sensitive to implicit comparison objects. For instance, when a talking toy robot was in the shape of a vehicle, people rated this robot as "extraverted" as they rated themselves. We think this happened because people were implicitly comparing the vehicle robot to other vehicles, which have no speech interactions at all with people.

In this study, we tested the sensitivity of our scales to this process by showing participants who rated Pearl one of three other objects to rate too—a car, a personal computer, and the Pyxis robot. We did not ask the participants to compare these objects, but nonetheless we found some sensitivity in ratings of the Pearl robot to the other object rated. The Pearl robot was rated as more efficient (F [2, 49] = 3.5, p < .05) and slightly more easily maintained (F [2, 49] = 3.0, p = .06) in comparison to the other robot, Pyxis, than in comparison either to a car or to a PC. However, the other scales (Durability, Safety, and Information Technology) were not significantly sensitive to the comparison objects.

Discussion

In this study, we developed five scales to be used in measuring people's mechanistic mental models of robots and computer-based technology. In the future, the addition of items and psychometric studies will add to their reliability and validity. However, we believe we have demonstrated the viability of such measurement. Even though two of the four rated objects were unfamiliar robots, we still obtained systematic differences in responses to the different machines.

References

Goetz, J. & Kiesler, S. Cooperation with a Robotic Assistant, in *CHI '02 Extended Abstracts* (Minneapolis, MN, April 2002), ACM Press.

John, O. Donahue, E. & Kentle, R (1991). The Big Five Inventory - Versions 4a and 54. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Kiesler, S. & Goetz, J. Mental Models of Robotic Assistants, in *CHI '02 Extended Abstracts* (Minneapolis, MN, April 2002), ACM Press.

Warner, R.M., & Sugarman, D.B. (1986). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality & Social Psychology, 50*, 792-799.

Woodson, W. E., Tillman, B., & Tillman, P. (1992). Human factors design handbook. 2nd Edition, NY: McGraw Hill.

Table 1. Traits rated as Machinelike

Trait	Average Machinelike Rating (n=20)	Average Humanlike Rating (n=20)	T-test p-value
Complex	4.33	3.53	0.05
Specialized	3.80	2.93	0.05
Hazardous	3.00	2.13	0.05
Low maintenance	3.13	2.00	0.05
Easy to manipulate	3.73	2.80	0.01
Handy	3.93	3.00	0.01
Productive	4.30	3.27	0.01
Powerful	4.40	3.33	0.01
Effortless	3.00	1.87	0.01
Interactive	4.00	2.27	0.01
Informative	4.01	2.80	0.001
Quick	4.40	3.13	0.001
Safe	3.40	2.07	0.001
Sturdy	3.53	2.13	0.001
Efficient	4.40	3.00	0.001
Controllable	3.93	2.33	0.001
Durable	3.73	2.13	0.001
High quality	3.87	2.07	0.001
Repetitive	3.93	2.20	0.0001
Requires effort	4.07	2.33	0.0001
Out of date	3.87	1.80	0.0001
Accurate	4.60	2.40	0.0001
Routinized	4.27	2.07	0.0001
Cost efficient	3.73	1.53	0.0001
Precise	4.73	2.47	0.0001
Requires maintenance	4.53	2.20	0.0001
Heavy duty	3.87	1.53	0.0001
User friendly	3.87	1.53	0.0001
Portable	3.80	1.40	0.0001
Has a lot of features	4.53	2.13	0.0001
Can save time	4.20	1.67	0.0001
Breakable	4.70	2.13	0.0001
Could be improved	4.47	1.87	0.0001