Measuring the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots.

Christoph Bartneck
Department of Industrial Design
Eindhoven University of Technology
Den Dolech 2, 5600MB Eindhoven
The Netherlands
Phone +31 40 247 5175
c.bartneck@tue.nl

Dana Kulic
Nakamura & Yamane Lab
Department of Mechano-Informatics
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-8656, Japan
dana@ynl.t.u-tokyo.ac.jp

Elizabeth Croft
Department of Mechanical Engineering
University of British Columbia
6250 Applied Science Lane
Room 2054, Vancouver
Canada V6T 1Z4
ecroft@mech.ubc.ca

ABSTRACT

This study emphasizes the need for standardized measurement tools for human robot interaction (HRI). If we are to make progress in this field then we must be able to compare the results from different studies. A literature review has been performed on the measurements of five key concepts in HRI: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. The results have been distilled into five consistent questionnaires using semantic differential scales. We report reliability and validity indicators based on several empirical studies that used these questionnaires. It is our hope that these questionnaires can be used by robot developers to monitor their progress. Psychologists are invited to further develop the questionnaires by adding new concepts, and to conduct further validations where it appears necessary.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Evaluation/methodology

General Terms

Measurement, Human Factors, Standardization

Keywords

Human factors, robot, perception, measurement.

1. INTRODUCTION

The success of service robots and, in particular, of entertainment robots cannot be assessed only by performance criteria typically found for industrial robots. The number of processed pieces and their accordance with quality standards are not necessarily the prime objectives for an entertainment robot such as Aibo (Sony, 1999), or a communication platform such as iCat (Breemen, Yan, & Meerbeek, 2005). The performance criteria of service robots lie within the satisfaction of their users. Therefore, it is necessary to measure the users' perception of service robots, since these can not be measured within the robots themselves.

Measuring human perception and cognition has its own pitfalls, and psychologists have developed extensive methodologies and statistical tests to objectify the acquired data. Most engineers who develop robots are often unaware of this large body of knowledge, and sometimes run naïve experiments in order to verify their designs. But the same naivety can also be expected of psychologists when confronted with the task of building a robot. Human-Robot Interaction (HRI) is a multidisciplinary field, but it can not be expected that everyone masters all skills equally well. We do not intend to investigate the structure of the HRI community and the problems it is facing in the cooperation of its members. The interested reader may consult Bartneck & Rauterberg (Bartneck & Rauterberg, 2007) who reflected on the structure of the Human-Computer Interaction community This may also apply to the HRI community. This study is intended for the technical developers of interactive robots who want to evaluate their creations without having to take a degree in experimental psychology. However, it is advisable to at least consult with a psychologist over the overall methodology of the experiment.

A typical pitfall in the measurement of psychological concepts is to break them down into smaller, presumably better-known, components. This is common practice, and we do not intend to single out a particular author, but we still feel the need to present an example. Kiesler and Goetz (2002) divided the concept of anthropomorphism into the sub components sociability, intellect, and personality. They measured each concept with the help of a questionnaire. This breaking down into sub components makes sense if the relationship and relative importance of the sub components are known and can therefore be calculated back into the original concept. Otherwise, a presumably vague concept is simply replaced by series of just as vague concepts. There is no reason to believe that it would be easier for the users of robots to evaluate their sociability rather than their anthropomorphism. Caution is therefore necessary so as not to over-decompose concepts. Still, it is good practice to at least decompose the concept under investigation into several items[1] so as to have richer and more reliable data as was suggested by Fink, volume 8, p. 20 (2003).

A much more reliable and possibly objective method for measuring the users' perception and cognition is to observe their behavior. If, for example, the intention of a certain robot is to play a game with the user, then the fun experienced can be deduced from the time the user spends playing it. The longer the user plays, the more fun it is. However, not all internal states of a user manifest themselves in observable behavior. From a practical point of view it can also be very laborious to score the users' behaviors on the basis of video recordings.

Physiological measurements form a second group of measurement tools. Skin conductivity, heart rate, and heart variance are three popular measurements that provide a good indication of the user's

[1] In the social sciences the term "item" refers to a single question or response.

arousal in real time. The measurement can be taken during the interaction with the robot. Unfortunately, these measurements can not distinguish the arousal that stems from anger from that which may originate from joy. To gain better insight into the user's state, these measurements can be complemented by other physiological measurements, such as the recognition of facial expression. In combination, they can provide real time data, but the effort of setting up and maintaining the equipment and software should not be underestimated.

A third measurement technique is questionnaires, which are often used to measure the users' attitudes. While this method is rather quick to conduct, its conceptual pitfalls are often underestimated. One of its prime limitations is, of course, that the questionnaire can be administered only after the actual experience. Subjects have to reflect on their experience afterwards, which might bias their response. They could, for example, adapt their response to the socially acceptable response.

The development of a validated questionnaire involves a considerable amount of work, and extensive guidelines are available to help with the process (Dawis, 1987; Fink, 2003). Development will typically begin with a large number of items, which are intended to cover the different facets of the theoretical construct to be measured; next, empirical data is collected from a sample of the population to which the measurement is to be applied. After appropriate analysis of this data, a subset of the original list of items is then selected and becomes the actual multi-indicator measurement. This measurement will then be formally assessed with regard to its reliability, dimensionality, and validity.

Due to their naivety and the amount of work necessary to create a validated questionnaire, developers of robots have a tendency to quickly cook up their own questionnaires. This conduct results in two main problems. Firstly, the validity and reliability of these questionnaires has often not been evaluated. An engineer is unlikely to trust a voltmeter developed by a psychologist unless its proper function has been shown. In the same manner, psychologists will have little trust in the results from a questionnaire developed by an engineer unless information about its validity and reliability is available. Secondly, the absence of standard questionnaires makes it difficult to compare the results from different researchers. If we are to make progress in the field of human-robot interaction then we shall have to develop standardized measurement tools similar to the ITC-SOPI questionnaire that was developed to measure presence (Lessiter, Freeman, Keogh, & Davidoff, 2001).

This study attempts to make a start in the development of standardized measurement tools for human-robot interaction by first presenting a literature review on existing questionnaires, and then presenting empirical studies that give an indication of the validity and reliability of these new questionnaires. This study will take the often-used concepts of anthropomorphism, animacy, likeability, and perceived intelligence and perceived safety as starting points to propose a consistent set of five questionnaires for these concepts.

We can not offer an exhaustive framework for the perception of robots similar to the frameworks that have already been developed for social robots (Bartneck & Forlizzi, 2004; Fong, Nourbakhsh, & Dautenhahn, 2003) that would justify the selection of these five concepts. We can only hint at the fact that the concepts proposed have been necessary for our own research and that they are likely to have relationships with each other. A highly anthropomorphic and intelligent robot is likely to be perceived to be more animate and possibly also more likeable. The verification of such a model does require appropriate measurement instruments. The discussion of whether it is good practice to first develop a theory and then the observation method or vice versa has not reached a conclusion (Chalmers, 1999), but every journey begins with a first step. The proposed set of questionnaires can later be extended to cover other relevant concepts, and their relationships can be further explored. The emphasis is on presenting questionnaires that can be used directly in the development of interactive robots. Many robots are being built right now, and the engineers cannot wait for a mature model to emerge. We even seriously consider the position that such a framework can be created only once we have the robots and measurement tools in place.

Unfortunately, the literature review revealed questionnaires that used different types of items, namely Likert-scales (Likert, 1932) and semantic differential scales (Osgood, Suci, & Tannenbaum, 1957). If more than one questionnaire is to be used for the evaluation of a certain robot, it is beneficial if the questionnaires use the same type of items. This consistency makes it easy for the participants to learn the method and thereby avoids errors in their responses. It was therefore decided to transfer Likert type scales to semantic differential scales. We shall now discuss briefly the differences between these two types of items.

In semantic differential scales the respondent is asked to indicate his or her position on a scale between two bipolar words, the anchors (see Figure 1, top). In Likert scales (see Figure 1, bottom), subjects are asked to respond to a stem, often in the form of a statement, such as "I like ice cream". The scale is frequently anchored with choices of "agree" - "disagree" or "like" - "dislike".

Strong 1 2 3 4 5 Weak

I like ice cream Disagree 1 2 3 4 5 Agree

Figure 1. Example of a semantic differential scale (top) and likert scale (bottom). The participant would be asked to rate the stimulus on this scale by circling one of the numbers.

Both are rating scales, and provided that response distributions are not forced, semantic differential data can be treated just as any other rating data (Dawis, 1987). The statistical analysis is identical. However, a semantic differential format may effectively reduce acquiescence bias without lowering psychometric quality (Friborg, Martinussen, & Rosenvinge, 2006). A common objection to Osgood's semantic differential method is that it appears to assume that the adjectives chosen as anchors mean the same to everyone. Thus, the method becomes self-contradictory; it starts from the presumption that different people interpret the same word differently, but has to rely on the assumption that this is not true for the anchors. However, this study proposes to use the semantic differential scales to evaluate not the meaning of words, but the attitude towards robots. Powers and Kiesler (2006) report a negative correlation (-.23) between Humanlikeness and Machinelikeness, which strengthens our view that semantic differentials are a useful tool for measuring the users' perception of robots, while we remain aware of the fact that every method has its limitations.

Some information on the validity and reliability of the questionnaires is already available from the original studies on which they are based. However, the transformation from Likert scales to semantic differential scales may compromise these indicators to a certain degree. We shall compensate this possible loss by reporting on complementary empirical studies later in the text. First, we would like to discuss the different types of validity and reliability.

Fink in Volume 8, pp 5-44, (Fink, 2003) discusses several forms of reliability and validity. Among the scientific forms of validity we find content validity, criterion validity, and construct validity. The latter, which determines the degree to which the instrument works in comparison with others, can only be assessed after years of experience with a questionnaire, and construct validity is often not calculated as a quantifiable statistic. Given the short history of research in HRI it would appear difficult to achieve construct validity. The same holds true for criterion validity. There is a scarcity of validated questionnaires with which our proposed questionnaires can be compared. We can make an argument for content validity since experts in the field carried out the original studies, and measurements of the validity and reliability have even been published from time to time. The researchers involved in the transformation of the proposed questionnaires were also in close contact with relevant experts in the field with regard to the questionnaires. The proposed questionnaires can therefore be considered to have content validity.

It is easier to evaluate the reliability of the questionnaire, and Fink describes three forms: test-retest reliability, alternate form reliability, and internal consistency reliability. The latter is a measurement for how well the different items measure the same concept, and it is of particular importance to the questionnaires proposed because they are designed to be homogenous in content. Internal consistency involves the calculation of a statistic known as Cronbach’s Alpha. It measures the internal consistency reliability among a group of items that are combined to form a single scale. It reflects the homogeneity of the scale. Given the choice of homogeneous semantic differential scales, alternate form reliability appears difficult to achieve. The items cannot simply be negated and asked again because semantic differential scales already include dichotomous pairs of adjectives. Test-retest reliability can even be tested within the same experiment by splitting the participants randomly into two groups. This procedure requires a sufficiently large number of participants and unfortunately none of the studies that we have access to had enough participants to allow for a meaningful test-retest analysis. For both, test-retest reliability and internal consistency reliability, Nunnally (1978) recommends a minimum value of 0.7. We would now like to discuss the five concepts of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety in more detail, and describe a questionnaire for each of them.

2. ANTHROPOMORPHISM

Anthropomorphism refers to the attribution of a human form, human characteristics, or human behavior to nonhuman things such as robots, computers, and animals. Hiroshi Ishiguro, for example, develops androids that, for a short period, are indistinguishable from human beings (Ishiguro, 2005). His highly anthropomorphic androids struggle with the so-called ‘uncanny valley’, a theory that states that as a robot is made more humanlike in its appearance and movements, the emotional response from a human being to the robot becomes increasingly positive and empathic, until a point is reached beyond which the response quickly becomes that of intense repulsion. However, as the appearance and movements continue to become less distinguishable from those of a human being, the emotional response becomes positive once more and approaches human-human empathy levels.

Even if it is not the intention of the design of a certain robot to be as humanlike as possible, it still remains important to match the appearance of the robot with its abilities. A too anthropomorphic appearance can evoke expectations that the robot might not be able to fulfill. If, for example, the robot has a human-shaped face then the naïve user will expect that the robot is able to listen and to talk. To prevent disappointment it is necessary for all developers to pay close attention to the anthropomorphism level of their robots.

An interesting behavioral measurement for anthropomorphism has been presented by Minato et al. (2005). They attempted to analyze differences in where the participants were looking when they looked at either a human or an android. The hypothesis is that people look differently at humans compared to robots. They have not been able to produce reliable conclusions yet, but their approach could turn out to be very useful, assuming that they can overcome the technical difficulties.

MacDorman (2006) presents an example of a naïve questionnaire. A single question is asked to assess the human-likeness of what is being viewed (9-point semantic differential, mechanical versus humanlike). It is good practice in the social sciences to ask multiple questions about the same concept in order to be able to check the participants’ consistency and the questionnaire’s reliability. Powers and Kiesler (2006), in comparison, used six items and are able to report a Cronbach’s Alpha of 0.85. Their questionnaire therefore appears to be more suitable. It was necessary to transform the items used by Powers and Kiesler into semantic differentials: Fake / Natural, Machinelike / Humanlike, Unconscious / Conscious, Artificial / Lifelike, and Moving rigidly / Moving elegantly.

Two studies are available in which this new anthropomorphism questionnaire was used. The first one reports a Cronbach’s Alpha of 0.878 (Bartneck, Kanda, Ishiguro, & Hagita, 2007) and we would like to report the Cronbach’s Alphas for the second study (Bartneck, Kanda, Ishiguro, & Hagita, 2008) in this paper. The study consisted of three within conditions for which the Cronbach’s Alphas must be reported separately. We can report a Cronbach’s Alpha of 0.929 for the human condition, 0.923 for the android condition and 0.856 for the masked android condition. The alpha values are well above 0.7, so we can conclude that the anthropomorphism questionnaire has sufficient internal consistency reliability.

3. ANIMACY

The goal of many robotics researchers is to make their robots lifelike. Computer games, such as The Sims, Creatures, or Nintendo Dogs show that lifelike creatures can deeply involve users emotionally. This involvement can then be used to influence users (Fogg, 2003). Since Heider and Simmel (1944), a considerable amount of research has been devoted to the perceived animacy and “intentions” of geometric shapes on computer screens. Scholl and Tremoulet (2000) offer a good summary of the research field, but, on examining the list of references, it becomes apparent that only two of the 79 references deal directly with animacy. Most of the reviewed work focuses on causality and intention. This may indicate that the measurement of animacy is difficult. Tremoulet and Feldman (2000) only asked their participants to evaluate the animacy of ‘particles’ under a microscope on a single scale (7-point Likert scale, 1=definitely not alive, 7=definitely alive). It is questionable how much sense it makes to ask participants about the animacy of particles. By definition they cannot be alive since particles tend to be even smaller than the simplest organisms.

Asking about the perceived animacy of a certain stimulus makes sense only if there is a possibility for it to be alive. Robots can show physical behavior, reactions to stimuli, and even language skills. These are typically attributed only to animals, and hence it can be argued that it makes sense to ask participants about their perception of the animacy of robots.

McAleer, et al. (2004) claim to have analyzed the perceived animacy of modern dancers and their abstractions on a computer screen, but only qualitative data of the perceived arousal is presented. Animacy was measured with free responses. They looked for terms and statements that indicated that subjects had attributed human movements and characteristics to the shapes. These were terms such as “touched”, “chased”, and “followed”, and emotions such as “happy” or “angry”. Other guides to animacy were when the shapes were generally being described in active roles, as opposed to being controlled in a passive role. However, they do not present any quantitative data for their analysis.

A better approach has been presented by Lee, Kwan Min, Park, Namkee & Song, Hayeon (2005). With their four items (10-point Likert scale; lifelike, machine-like, interactive, responsive) they have been able to achieve a Cronbach’s Alpha of 0.76. For the questionnaires in this study, their items have been transformed into semantic differentials: Dead / Alive, Stagnant / Lively, Mechanical / Organic, Artificial / Lifelike, Inert / Interactive, Apathetic / Responsive. One study used this new questionnaire (Bartneck, Kanda, Mubin, & Mahmud, 2007) and reported a Cronbach’s Alpha of 0.702, which is sufficiently high for us to conclude that the new animacy questionnaire has sufficient internal consistency reliability.

4. LIKEABILITY

It has been reported that the way in which people form positive impressions of others is to some degree dependent on the visual and vocal behavior of the targets (Clark & Rutter, 1985), and that positive first impressions (e.g., likeability) of a person often lead to more positive evaluations of that person (Robbins & DeNisi, 1994). Interviewers report knowing within 1 to 2 minutes whether a potential job applicant is a winner, and people report knowing within the first 30 seconds the likelihood that a blind date will be a success (Berg & Piner, 1990). There is a growing body of research indicating that people often make important judgments within seconds of meeting a person, sometimes remaining quite unaware of both the obvious and subtle cues that may be influencing their judgments. Since computers, and thereby robots in particular, are to some degree treated as social actors (Nass & Reeves, 1996), it can be assumed that people are able to judge robots just as.

Jennifer Monathan (1998) complemented her “liking” question with 5-point semantic differential scales: nice / awful, friendly / unfriendly, kind / unkind, and pleasant / unpleasant, because these judgments tend to demonstrate considerable variance in common with “liking” judgments (Burgoon & Hale, 1987). Monahan later eliminated the kind-unkind and pleasant-unpleasant items in her own analysis since they did not load sufficiently in a factor analysis that also included items from three other factors. The Cronbach’s Alpha of 0.68 therefore relates only to this reduced scale. Her experimental focus is different from the intended use of her questionnaire in the field of HRI. She also included concepts of physical attraction, conversational skills, and other orientations, which might not be of prime relevance to HRI. In particular, physical attraction might be unsuitable for robots. No reports on successful human-robot reproduction are available yet and hopefully never will be. We decided to only include the five items, since it is always possible to exclude items in cases where they would not contribute to the reliability and validity of the questionnaire.

Two studies used this new likeability questionnaire. The first reports a Cronbach’s Alpha of 0.865 (Bartneck, Kanda, Ishiguro, & Hagita, 2007), and we report the Cronbach’s Alpha for the second (Bartneck, Kanda, Ishiguro, & Hagita, 2008) in this paper. The study consisted of three “within” conditions for which the Cronbach’s Alpha must be reported separately. Without going into too much detail of the study, we can report a Cronbach’s Alpha of 0.923 for the human condition, 0.878 for the android condition, and 0.842 for the masked android condition. The alpha values are well above 0.7, and hence we can conclude that the likeability questionnaire has sufficient internal consistency reliability.

5. PERCEIVED INTELLIGENCE

Interactive robots face a tremendous challenge in acting intelligently. The reasons can be traced back to the field of artificial intelligence (AI). The robots’ behaviors are based on methods and knowledge that were developed by AI. Many of the past promises of AI have not been fulfilled, and AI has been criticized extensively (Dreyfus & Dreyfus, 1992; Dreyfus, Dreyfus, & Athanasiou, 1986; Searle, 1980; Weizenbaum, 1976).

One of the main problems that AI is struggling with is the difficulty of formalizing human behavior, for example, in expert systems. Computers require this formalization to generate intelligent and human-like behavior. And as long as the field of AI has not made considerable progress on these issues, robot intelligence will remain at a very limited level. So far, we have been using many Wizard-Of-Oz methods to fake intelligent robotic behavior, but this is possible only in the confines of the research environment. Once the robots are deployed in the complex world of everyday users, their limitations will become apparent. Moreover, when the users are interacting with the robot for years rather than minutes, they will become aware of the limited abilities of most robots.

Evasion strategies have also been utilized. The robot would show more or less random behavior while interacting with the user, and the user in turn sees patterns in this behavior which he/she interprets as intelligence. Such a strategy will not lead to a solution of the problem, and its success is limited to short interactions. Given sufficient time the user will give up his/her hypothesized patterns of the robot’s intelligent behavior and become bored with its limited random vocabulary of behaviors. In the end, the perceived intelligence of a robot will depend on its competence (Koda, 1996). To monitor the progress being made in robotic intelligence it is important to have a good measurement tool.

Warner and Sugarman (1996) developed an intellectual evaluation scale that consists of five seven-point semantic differential items: Incompetent / Competent, Ignorant / Knowledgeable, Irresponsible / Responsible, Unintelligent / Intelligent, Foolish / Sensible. Parise et al. (Parise, Kiesler, Sproull , & Waters 1996) excluded one question from this scale, and reported a Cronbach's Alpha of 0.92. The questionnaire was again used by Kiesler, Sproull and Waters (Kiesler, Sproull, & Waters, 1996), but no alpha was reported. Three other studies used the perceived intelligence questionnaire, and reported Cronbach's Alpha values of 0.75 (Bartneck, Kanda, Ishiguro, & Hagita, 2008), 0.769 (Bartneck, Verbunt, Mubin, & Mahmud, 2007), and 0.763 (Bartneck, Kanda, Mubin, & Mahmud, 2007). These values are above the suggested 0.7 threshold, and hence the perceived intelligence questionnaire can be considered to have satisfactory internal consistency reliability.

6. PERCEIVED SAFETY

Perceived safety describes the user's perception of the level of danger when interacting with a robot, and the user's level of comfort during the interaction. Achieving a positive perception of safety is a key requirement if robots are to be accepted as partners and co-workers in human environments. Perceived safety and user comfort have rarely been measured directly. Instead, indirect measures have been used - the measurement of the affective state of the user through the use of physiological sensors (Kulic & Croft, 2005; Rani, Sarkar, Smith, & Kirby, 2004; Rani, Sims, Brackin, & Sarkar, 2002), questionnaires (Inoue, Nonaka, Ujiie, Takubo, & Arai, 2005; Kulic & Croft, 2005; Wada, Shibata, Saito, & Tanie, 2004), and direct input devices (Koay, Walters, & Dautenhahn, 2005). That is, instead of asking subjects to evaluate the robot, researchers frequently use affective state estimation or questionnaires asking how the subject feels in order to measure the perceived safety and comfort level indirectly.

For example, Sarkar proposes the use of multiple physiological signals to estimate affective state, and to use this estimate to modify robotic actions to make the user more comfortable (Sarkar, 2002). Rani et al. (2004; 2002) use heart-rate analysis and multiple physiological signals to estimate human stress levels. In Rani et al. (2004), an autonomous mobile robot monitors the stress level of the user, and if the level exceeds a certain value, the robot returns the user in a simulated rescue attempt. However, in their study, the robot does not interact directly with the human; instead, pre-recorded physiological information is used to allow the robot to assess the human's condition.

Koay et al. (2005) describe an early study where human reaction to robot motions was measured online. In this study, 28 subjects interacted with a robot in a simulated living room environment. The robot motion was controlled by the experimenters in a "Wizard of Oz" setup. The subjects were asked to indicate their level of comfort with the robot by means of a handheld device. The device consisted of a single slider control to indicate comfort level, and a radio signal data link. Data from only 7 subjects was considered reliable, and was included in subsequent analysis. Analysis of the device data with the video of the experiment found that subjects indicated discomfort when the robot was blocking their path, the robot was moving behind them, or the robot was on a collision course with them.

Nonaka et al (2004) describe a set of experiments where human response to pick-and-place motions of a virtual humanoid robot is evaluated. In their experiment, a virtual reality display is used to depict the robot. Human response is measured through heart rate measurements and subjective responses. A 6-level scale is used from 1 = "never" to 6 = "very much", for the categories of "surprise", "fear", "disgust", and "unpleasantness". No relationship was found between the heart rate and robot motion, but a correlation was reported between the robot velocity and the subject's rating of "fear" and "surprise". In a subsequent study (Inoue, Nonaka, Ujiie, Takubo, & Arai, 2005), a physical mobile manipulator was used to validate the results obtained with the virtual robot. In this case, subjects are asked to rate their responses on the following (5-point) direction levels: "secure – anxious", "restless – calm", "comfortable – unpleasant", "unapproachable – accessible", "favorable – unfavorable", "tense – relaxed", "unfriendly – friendly", "interesting – tedious", and "unreliable – reliable". They are also asked to rate their level of "intimidated" and "surprised" on a 5 –point Likert scale. The study finds that similar results are obtained regardless of whether a physical or a virtual robot is used. Unfortunately, no information about the reliability or validity of their scales is available. There is a very large number of different questions that can be asked on the topic of safety and comfort in response to physical robot motion. This underlines the need for a careful and studied set of baseline questions for eliciting comparable results from research efforts, especially in concert with physiological measurement tools. It becomes apparent that two approaches can be taken to assess the perceived safety. On the one hand the users can be asked to evaluate their impression of the robot, and on the other hand they can be asked to assess their own affective state. It is assumed that if the robot is perceived to be dangerous then the user affective state would be tense.

Kulic and Croft (2005) combined a questionnaire with physiological sensors to estimate the user's level of anxiety and surprise during sample interactions with an industrial robot. They ask the user to rate their level of anxiety, surprise, and calmness during each sample robot motion. A 5 point Likert scale is used. The Cronbach's Alpha for the affective state portion of the questionnaire is 0.91. In addition, the subject is asked to rate their level of attention during the robot motion, to ensure that the elicited affective state was caused by the robot rather than by some other internal or external distraction. In this work, they show that motion planning can be used to reduce the perceived anxiety and surprise felt by subjects during high speed movements. This and later work (Kulic & Croft, 2006) by the same authors showed a strong statistical correlation between the affective state reported by the subjects and their physiological responses. The scales they produced can be transformed to the following semantic differential scales: Anxious / Relaxed, Agitated / Calm, Quiescent / Surprised. This questionnaire focuses on the affective state of the user. To our knowledge, no suitable questionnaire for rating the safety of a robot is available.

7. CONCLUSIONS

The study proposes a series of questionnaires to measure the users' perception of robots. This series will be called "Godspeed" because it is intended to help creators of robots on their development journey. Appendix A shows the application of the five Godspeed questionnaires using 5-point scales. It is important to notice that there is a certain overlap between anthropomorphism and animacy. The item artificial / lifelike appears in both sections. This is to be expected, since being alive is an essential part of being human-like.

When one of these questionnaires is used by itself in a study it would be useful to mask the questionnaire's intention by adding dummy items, such as optimistic / pessimistic. If multiple questionnaires are used then the items should be mixed so as to mask the intention. Before calculating the mean scores for anthropomorphism, animacy, likeability, or perceived intelligence it is good practice to perform a reliability test and report the resulting Cronbach's Alpha.

The interpretation of the results has, of course, some limitations. First, it is extremely difficult to determine the ground truth. In other words, it is complicated to determine objectively, for example, how anthropomorphic a certain robot is. Many factors, such as the cultural backgrounds of the participants, prior experiences with robots, and personality may influence the measurements. Taking all the possible biases into account would require a complex and therefore impracticable experiment. The resulting values of the measurements should therefore be interpreted not as absolute values, but rather as a tool for comparison. Robot developers can, for example, use the questionnaires to compare different configurations of a robot. The results may then help the developers to choose one option over the other. In the future, this set of questionnaires could be extended to also include the believability of a robot, the enjoyment of interacting with it, and the robot's social presence.

It is the hope of the authors that robot developers may find this collection of measurement tools useful. Using these tools would make the results in HRI research more comparable and could therefore increase our progress. Interested readers, in particular experimental psychologists, are invited to continue to develop these questionnaires, and to validate them further.

A necessary development would be translation into different languages. Only native speakers can understand the true meanings of the adjectives in their language. It is therefore necessary to translate the questionnaires into the mother language of the participants. Appendix A includes the Japanese translation of the adjectives that we created using the back translation method. It is advisable to use the same method to translate the questionnaire into other languages. It would be appreciated if other translations are reported back to the authors of this study. They will then be collected and posted on this website:

http://www.bartneck.de/work/researchProjects/socialRobotics/godspeed

8. ACKNOWLEDGMENTS

The Intelligent Robotics and Communication Laboratories at the Advanced Telecommunications Institute International (Kyoto, Japan) supported this study.

9. REFERENCES

Bartneck, C., & Forlizzi, J. (2004). *A Design-Centred Framework for Social Human-Robot Interaction.* Proceedings of the Ro-Man2004, Kurashiki pp. 591-594. | DOI: 10.1109/ROMAN.2004.1374827

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2007). *Is the Uncanny Valley an Uncanny Cliff?* Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2007, Jeju, Korea pp. 368-373. | DOI: 10.1109/ROMAN.2007.4415111

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2008). My Robotic Doppelgänger – A Critical Look at the Uncanny Valley Theory. *Autonomous Robots*.

Bartneck, C., Kanda, T., Mubin, O., & Mahmud, A. A. (2007). *The Perception of Animacy and Intelligence Based on a Robot's Embodiment.* Proceedings of the Humanoids 2007, Pittsburgh.

Bartneck, C., & Rauterberg, M. (2007). HCI Reality - An Unreal Tournament. *International Journal of Human Computer Studies, 65*(8), 737-743. | DOI: 10.1016/j.ijhcs.2007.03.003

Bartneck, C., Verbunt, M., Mubin, O., & Mahmud, A. A. (2007). *To kill a mockingbird robot.* Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction, Washington DC pp. 81-87. | DOI: 10.1145/1228716.1228728

Berg, J. H., & Piner, K. (1990). Social relationships and the lack of social relationship. In W. Duck & R. C. Silver (Eds.), *Personal relationships and social support* (pp. 104-221). London: Sage.

Breemen, A., Yan, X., & Meerbeek, B. (2005). *iCat: an animated user-interface robot with personality.* Proceedings of the Fourth International Conference on Autonomous Agents & Multi Agent Systems, Utrecht. | DOI: 10.1145/1082473.1082823

Burgoon, J. K., & Hale, J. L. (1987). Validation and measurement of the fundamental themes for relational communication. *Communication Monographs, 54*, 19-41.

Chalmers, A. F. (1999). *What is this thing called science?* (3rd ed.). Indianapolis: Hackett.

Clark, N., & Rutter, D. (1985). Social categorization, visual cues and social judgments. *European Journal of Social Psychology, 15*, 105-119. | DOI: 10.1002/ejsp.2420150108

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489. | DOI: 10.1037/0022-0167.34.4.481

Dreyfus, H. L., & Dreyfus, S. E. (1992). *What computers still can't do : a critique of artificial reason*. Cambridge, Mass.: MIT Press.

Dreyfus, H. L., Dreyfus, S. E., & Athanasiou, T. (1986). *Mind over machine : the power of human intuition and expertise in the era of the computer*. New York: Free Press.

Fink, A. (2003). *The survey kit* (2nd ed.). Thousand Oaks, Calif.: Sage Publications.

Fogg, B. J. (2003). *Persuasive technology : using computers to change what we think and do*. Amsterdam ; Boston: Morgan Kaufmann Publishers.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*, 143-166. | DOI: 10.1016/S0921-8890(02)00372-X

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences, 40*(5), 873-884. | DOI: 10.1016/j.paid.2005.08.015

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57,* 243-249.

Inoue, K., Nonaka, S., Ujiie, Y., Takubo, T., & Arai, T. (2005). *Comparison of human psychology for real and virtual mobile manipulators.* Proceedings, IEEE International Conference on

Robot and Human Interactive Communication pp. 73 - 78. | DOI: 10.1109/ROMAN.2005.1513759

Ishiguro, H. (2005). *Android Science - Towards a new cross-interdisciplinary framework.* Proceedings of the CogSci Workshop Towards social Mechanisms of android science, Stresa pp. 1-6.

Kiesler, S., & Goetz, J. (2002). *Mental models of robotic assistants.* Proceedings of the CHI '02 extended abstracts on Human factors in computing systems, Minneapolis, Minnesota, USA. | DOI: 10.1145/506443.506491

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of personality and social psychology 70*(1), 47-65. | DOI: 10.1037/0022-3514.70.1.47

Koay, K. L., Walters, M. L., & Dautenhahn, K. (2005). *Methodological Issues Using a Comfort Level Device in Human-Robot Interactions.* Proceedings of the IEEE RO-MAN pp. 359 - 364.

Koda, T. (1996). *Agents with Faces: A Study on the Effect of Personification of Software Agents.* Master Thesis, MIT Media Lab, Cambridge.

Kulic, D., & Croft, E. (2005). *Anxiety Detection during Human-Robot Interaction.* Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Edmonton, Canada pp. 389 - 394. | DOI: 10.1109/IROS.2005.1545012

Kulic, D., & Croft, E. (2006). *Estimating Robot Induced Affective State Using Hidden Markov Models.* Proceedings of the RO-MAN 2006 – The 15th IEEE International Symposium on Robot and Human Interactive Communication, Hatfield pp. 257-262. | DOI: 10.1109/ROMAN.2006.314427

Lee, K. M., Park, N., & Song, H. (2005). Can a Robot Be Perceived as a Developing Creature? *Human Communication Research, 31*(4), 538-563. | DOI: 10.1111/j.1468-2958.2005.tb00882.x

Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J. (2001). A cross-media presence questionnaire: The itc sense of presence inventory. *Presence: Teleoperators and Vir tual Environments, 10*(3), 282-297. | DOI: 10.1162/105474601300343612

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*.

MacDorman, K. F. (2006). *Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley.* Proceedings of the ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science, Vancouver.

McAleer, P., Mazzarino, B., Volpe, G., Camurri, A., Patterson, H., & Pollick, F. (2004). Perceiving Animacy and Arousal in Transformed Displays of Human Interaction. *Journal of Vision, 4*(8), 230-230. | DOI: 10.1167/4.8.230

Minato, T., Shimada, M., Itakura, S., Lee, K., & Ishiguro, H. (2005). *Does Gaze Reveal the Human Likeness of an Android?* Proceedings of the 4th IEEE International Conference on Development and Learning, Osaka. | DOI: 10.1109/DEVLRN.2005.1490953

Monathan, J. L. (1998). I Don't Know It But I Like You - The Influence of Non-conscious Affect on Person Perception. *Human Communication Research, 24*(4), 480-500. | DOI: 10.1111/j.1468-2958.1998.tb00428.x

Nass, C., & Reeves, B. (1996). *The Media equation*. Cambridge: SLI Publications, Cambridge University Press.

Nonaka, S., Inoue, K., Arai, T., & Mae, Y. (2004). *Evaluation of Human Sense of Security for Coexisting Robots using Virtual Reality.* Proceedings of the IEEE International Conference on Robotics and Automation, New Orleans, LA, USA pp. 2770-2775. | DOI: 10.1109/ROBOT.2004.1307480

Nunnally, J. C. (1978). *Psychometric theory* (2d ed.). New York: McGraw-Hill.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurements of meaning*. Champaign: University of Illinois Press.

Parise, S., Kiesler, S., Sproull , L. D., & Waters , K. (1996). *My partner is a real dog: cooperation with social agents.* Proceedings of the 1996 ACM conference on Computer supported cooperative work, Boston, Massachusetts, United States pp. 399-408. | DOI: 10.1145/240080.240351

Powers, A., & Kiesler, S. (2006). *The advisor robot: tracing people's mental model from a robot's physical attributes.* Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, Salt Lake City, Utah, USA. | DOI: 10.1145/1121241.1121280

Rani, P., Sarkar, N., Smith, C. A., & Kirby, L. D. (2004). Anxiety detecting robotic system - towards implicit human-robot collaboration. *Robotica, 22*, 85-95. | DOI: 10.1017/S0263574703005319

Rani, P., Sims, J., Brackin, R., & Sarkar, N. (2002). Online stress detection using phychophysiological signals for implicit human-robot cooperation. *Robotica, 20*(6), 673-685. | DOI: 10.1017/S0263574702004484

Robbins, T., & DeNisi, A. (1994). A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology, 79*, 341-353. | DOI: 10.1037/0021-9010.79.3.341

Sarkar, N. (2002). *Psychophysiological Control Architecture for Human-Robot Coordination - Concepts and Initial Experiments.* Proceedings of the IEEE International Conference on Robotics and Automation, Washington, DC, USA pp. 3719-3724. | DOI: 10.1109/ROBOT.2002.1014287

Scholl, B., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*(8), 299-309. | DOI: 10.1016/S1364-6613(00)01506-0

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*(3), 417-457.

Sony. (1999). Aibo. Retrieved January, 1999, from http://www.aibo.com

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception, 29*(8), 943-951. | DOI: 10.1068/p3101

Wada, K., Shibata, T., Saito, T., & Tanie, K. (2004). Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proceedings of the IEEE, 92*(11), 1780-1788. | DOI: 10.1109/JPROC.2004.835378

Warner, R. M., & Sugarman, D. B. (1996). Attributes of Personality Based on Physical Appearance, Speech, and Handwriting. *Journal of Personality and Social Psychology, 50*(4), 792-799. | DOI: 10.1037/0022-3514.50.4.792

Weizenbaum, J. (1976). *Computer power and human reason : from judgment to calculation*. San Francisco: W. H. Freeman.

Appendix A: Overview of the Godspeed Questionnaire series using a 5-point scale.

GODSPEED I: ANTHROPOMORPHISM

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Fake 偽物のような	1	2	3	4	5	Natural 自然な
Machinelike 機械的	1	2	3	4	5	Humanlike 人間的
Unconscious 意識を持たない	1	2	3	4	5	Conscious 意識を持っている
Artificial 人工的	1	2	3	4	5	Lifelike 生物的
Moving rigidly ぎこちない動き	1	2	3	4	5	Moving elegantly洗練された動き

GODSPEED II: ANIMACY

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Dead 死んでいる	1	2	3	4	5	Alive 生きている
Stagnant 活気のない	1	2	3	4	5	Lively 生き生きとした
Mechanical 機械的な	1	2	3	4	5	Organic 有機的な
Artificial 人工的な	1	2	3	4	5	Lifelike 生物的な
Inert 不活発な	1	2	3	4	5	Interactive 対話的な
Apathetic 無関心な	1	2	3	4	5	Responsive 反応のある

GODSPEED III: LIKEABILITY

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Dislike 嫌い	1	2	3	4	5	Like 好き
Unfriendly 親しみにくい	1	2	3	4	5	Friendly 親しみやすい
Unkind 不親切な	1	2	3	4	5	Kind 親切な
Unpleasant 不愉快な	1	2	3	4	5	Pleasant 愉快な
Awful ひどい	1	2	3	4	5	Nice 良い

GODSPEED IV: PERCEIVED INTELLIGENCE

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Incompetent 無能な	1	2	3	4	5	Competent 有能な
Ignorant 無知な	1	2	3	4	5	Knowledgeable 物知りな
Irresponsible 無責任な	1	2	3	4	5	Responsible 責任のある
Unintelligent 知的でない,	1	2	3	4	5	Intelligent 知的な
Foolish 愚かな	1	2	3	4	5	Sensible 賢明な

GODSPEED V: PERCEIVED SAFETY

Please rate your emotional state on these scales:
以下のスケールに基づいてあなたの心の状態を評価してください。

Anxious 不安な	1	2	3	4	5	Relaxed 落ち着いた
Agitated 動揺している	1	2	3	4	5	Calm 冷静な
Quiescent 平穏な	1	2	3	4	5	Surprised 驚いた