HUMAN EMOTIONS TOWARD STIMULI IN THE UNCANNY VALLEY: LADDERING AND INDEX CONSTRUCTION

Chin-Chang Ho

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics and Computing,

Indiana University

March 2015

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Mark S. Pfaff, Ph.D., Chair

Alexander Fedorikhin, Ph.D.

Doctoral Committee

Edgar Huang, Ph.D.

July 14, 2014

Karl F. MacDorman, Ph.D.

DEDICATION

To my family

ACKNOWLEDGMENTS

It is a great pleasure to thank those who have made this dissertation possible. First, I would like to express my gratitude to my advisor, Dr. Karl F. MacDorman. Without his help and guidance, it is impossible to finish this dissertation. His expertise, patience, and thoughtfulness added considerably to my graduate experience. I appreciate his knowledge and skills in many areas. Second, I would like to thank other members of my committee, Dr. Mark Pfaff, Dr. Edgar Huang, and Dr. Alexander Fedorikhin for their insightful comments. Third, I would like to thank Himalaya Patel, Wade Mitchell, Ryan Sukale, and others in the Android Science Center. Without their contribution, I would not have had enough data to finish this thesis. Fourth, I give credit to my parents and friends. Without their love, encouragement, and emotional support, I would not have been able to complete this thesis. Last, but not least, I especially owe sincere and earnest thankfulness to my wife, Jessica, for her love and patience, even in my darkest moments. Finally, I would like to thank my first-born daughter, Catherine. You always give me the strength to get through this magnificent journey.

Chin-Chang Ho

HUMAN EMOTIONS TOWARD STIMULI IN THE UNCANNY VALLEY: LADDERING AND INDEX CONSTRUCTION

Human-looking computer interfaces, including humanoid robots and animated humans, may elicit in their users eerie feelings. This effect, often called the uncanny valley, emphasizes our heightened ability to distinguish between the human and merely humanlike using both perceptual and cognitive approaches. Although reactions to uncanny characters are captured more accurately with emotional descriptors (e.g., *eerie* and *creepy*) than with cognitive descriptors (e.g., *strange*), and although previous studies suggest the psychological processes underlying the uncanny valley are more perceptual and emotional than cognitive, the deep roots of the concept of humanness imply the application of category boundaries and cognitive dissonance in distinguishing among robots, androids, and humans. First, laddering interviews ($N = 30$) revealed firm boundaries among participants' concepts of animated, robotic, and human. Participants associated human traits like *soul*, *imperfect*, or *intended* exclusively with humans, and they simultaneously devalued the autonomous accomplishments of robots (e.g., *simple task*, *limited ability*, or *controlled*). Jerky movement and humanlike appearance were associated with robots, even though the presented robotic stimuli were humanlike. The facial expressions perceived in robots as *improper* were perceived in animated characters as *mismatched*. Second, association model testing indicated that the independent evaluation based on the developed indices is a viable quantitative technique for the laddering interview. Third, from the interviews several candidate items for the eeriness index were validated in a large representative survey ($N = 1{,}311$). The improved eeriness

index is nearly orthogonal to perceived humanness ($r = .04$). The improved indices facilitate plotting relations among rated characters of varying human likeness, enhancing perspectives on humanlike robot design and animation creation.

Mark S. Pfaff, Ph.D., Chair

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Problem Statement

Domestic robots (e.g., Roomba, Scooba, and Braava) are decreasing in price and becoming increasingly common in households. In 2012, about 3 million service robots for personal and domestic purposes were sold, 20% more than in 2011. The value of sales was predicted to increase to US$1.2 billion in 2013 (International Federation of Robotics, 2013). Meanwhile, socially assistive robots have demonstrated their ability to function in everyday life, from encouragement in performing rehabilitation exercises to social mediation and intimate companionship (Dautenhahn & Werry, 2002; Feil-Seifer, Skinner, & Matarić, 2007; Iwamura et al., 2011; Kozima, Nakagawa, & Yasuda, 2005; Kanda, Nishio, Ishiguro, & Hagita, 2009; Turkle, 2007; Wada et al., 2005). Android robots are simulating the form, motion quality, and contingent interaction of humans with ever more realism (Beck-Asano & Ishiguro, 2011; MacDorman et al., 2005; MacDorman & Ishiguro, 2006; MacDorman, 2006; Matsui, Minato, MacDorman, & Ishiguro, 2005; Sung, Guo, Grinter, & Christensen, 2007). Given the human desire for companionship and for nurturing others (Turkle, 2007), which is linked to our biological imperative, it is not hard to foresee the widespread use of humanlike robots once certain issues are resolved, such as cost of ownership and interaction difficulty.

Although robots have great potential to enhance daily life, people's attitudes toward robots strongly influence their acceptance of them. For example, negative attitudes and anxiety toward robots affects human emotional responses toward them and preferred distances to them (Nomura, Shintani, Fuji, & Hokabe, 2007). Human beings are highly sensitive to interpersonal responses and humanlike appearance because of

evolutionary selection and childhood learning (Rhodes & Zebrowitz, 2001). Only humanlike appearance and behavior can elicit fully humanlike communication (MacDorman & Ishiguro, 2006). However, Mori (1970/2012) cautioned against making robots that look too human because they could appear uncanny.[1] Powers, Kiesler, and Goetz (2003) showed people expected the performance of a robot to conform to expectations created by its appearance. Woods, Dautenhahn, and Schulz (2005) found that children and adults agree on the classifications of robot appearance, especially in machinelike and humanlike robots. However, children were more limited than adults in their ability to infer robot personalities and emotional states. Children might also have difficulties in initiating a relationship with robots. When the population includes not only children but older adults and people requiring medical care or assistance, designing social robots that interact well with these different populations will be a challenge.

An important issue in creating a design strategy for socially assistive robots and other anthropomorphic characters is how to measure human emotions while the participants are interacting with them. Without a validated evaluation, robot designers and computer animators choose oversimplified methods to evaluate their designs. The lack of a validated evaluation reduces the effectiveness of the design principles they develop for humanlike robots.

1.2 Past Work that Has Addressed the Problem

Little work has been done to create and validate measures of human emotion during interactions with humanlike robots or computer-generated (CG) characters.

[1] Film critics and computer graphics animators have also expressed such concerns in reference to the simulated human characters in films, such as *Polar Express* (2004) and *Final Fantasy: The Spirits Within* (2001; Butler & Joschko, 2007; Freedman, 2012; Plantec, 2007).

Bartneck, Kulić, Croft, and Zoghbi (2009) proposed the Godspeed indices based on such concepts as anthropomorphism, animacy, likability, perceived intelligence, and perceived safety, but these concepts often overlap. The negative attitude about robots scale (NARS; Nomura, Kanda, Suzuki, & Kato, 2004) evaluates the human rather than the human-robot interaction (Ho & MacDorman, 2010).

1.3 Questions Unanswered by Past Work

Robot designers and computer animators worry about their robot and computer graphics (CG) characters falling into the *uncanny valley* as they increase their human photorealism. Although designers and animators may debate whether their characters look "almost too real," the phenomenon of the uncanny valley is not yet well researched, especially when human emotions are involved. Most prior studies focus only on users' negative emotions or attitudes toward robots. Only a few comprehensive studies examined human emotions toward other humanlike entities (Ho & MacDorman, 2010). In addition, no validated indices exist for evaluating humanlike objects, which may otherwise guide the design of robots or CG characters. Instead, robot designers and computer animators routinely choose one of two ways to avoid falling into the uncanny valley: pushing realism to the practical limit or using a more abstract appearance. However, neither approach can solve the problem effectively, because humans may not take seriously a robot with a simplistic appearance, or they may be repulsed by a robot that looks human but still possesses subtle nonhuman features. One strategy to overcome these obstacles during the design process would be to systematically evaluate humanlike

characters and their interactions. Such a goal could be accomplished with a validated measure.

The mechanism of categorization, often found in cognitive psychology may provide a foundation for the development of such a measure. Ramey (2005, 2006) suggested that the uncanny valley reaction may be caused by objects that lie between categories rather than within them. Similar issues in cognitive psychology have been examined through categorization (e.g., the McGurk effect[2] in speech; McGurk & MacDonald, 1976). Could the discrimination among various humanlike entities be similar to the effect in color perception (e.g., the Sapir-Whorf hypothesis[3]) and between phoneme categories? The differences among various robots or among various androids might look much smaller than equal-sized differences across the robot–android boundary. The differences among humanlike entities show the category boundary is not merely quantitative but qualitative (Harnad, 1987). Only a few studies have examined the relation between the categories of humanlike objects and the uncanny valley effect (Ramey, 2005, 2006). The way we evaluate humanlike objects might be rooted in the mechanism of categorization. Therefore, exploring the categorization of robots can help us understand how people overestimate or underestimate robots.

---

[2] This perceptual phenomenon, discovered by Harry McGurk and John MacDonald, demonstrates an interaction between hearing and vision in speech perception. It occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound. Two common illusions in response to incongruent audiovisual stimuli have been observed: fusions ("ba-ba" auditory and "ga-ga" visual produce "da-da") and combinations ("ga-ga" auditory and "ba-ba" visual produce "bga-bga"). It shows that the brain's attempt to provide the consciousness with its best guess about the incoming information.

[3] It known as the linguistic relativity hypothesis, proposes a systematic relationship between the grammatical categories of the language a person speaks and how that person both understands the world and behaves in it.

1.4 Purpose of the Study

This study is intended to construct indices that measure the perceptions of anthropomorphic characters and to investigate the cognitive boundaries and dissonances among human-looking characters. A basic problem in the study of the uncanny valley is that the feelings of comfort or eeriness with a humanoid robot or CG character are strongly associated with human concepts (Becker-Asano, Ogawa, Nishio, & Ishiguro, 2010). Therefore, perceptions of human likeness are inevitably correlated with those of warmth and attractiveness. The indices of humanness and eeriness developed by Ho and MacDorman (2010) were designed to be decorrelated with warmth, and that result was confirmed. However, a humanlike appearance can cause users to over-interpret or otherwise misunderstand an agent's 'intentions' and actions. To solve this problem, this study took two steps. First, the laddering technique, a structured interview for uncovering core values, is used to determine the category boundary of an anthropomorphic character between human and robot. The terms gathered from the laddering interview as the candidates, can improve the indices of attractiveness, humanness, and eeriness. Second, the categorized responses will clarify category boundaries among humans, robots, and androids.

1.5 Significance of the Study

Robot designers routinely choose one of two ways to avoid falling into the uncanny valley. The first approach, pushing realism to the practical limit, can maximize our perception of human likeness in the robot. The second approach, using a more abstract appearance, helps eliminate aversion (DiSalvo, Gemperle, Forlizzi, & Kiesler, 2002). Before determining the guiding principles for a new robot, the designer should be

able to consult research to predict which emotions people will likely project onto the proposed robot. The trend toward creating robot companions could be jeopardized by a failure to take into account the role of appearance on user acceptance. Therefore, it is important to provide a framework to establish a comprehensive set of indices for humanlike entities. First, this work will improve previously constructed indices (Ho & MacDorman, 2010) to evaluate robots, androids, and anthropomorphic computer agents, to evaluate the goodness-of-fit of the indices, and to determine whether the indices are valid for the target objects. Second, the participant's categorizations of robot, animated character, and human can benefit our understanding of human perception. This work has explored how people understand the extent to which categories define humanlike objects. Third, the hierarchical value maps from the laddering interview should reveal the perceptual category boundaries, and the relation between perceptual attribution and human emotion. These improvements will facilitate the design of humanlike characters, our understanding of interactions with these characters, and ultimately, our acceptance of these characters.

2 LITERATURE REVIEW

2.1 The Uncanny Valley

One of the critical issues in human–robot interaction (HRI) is the uncanny valley (*bukimi no tani* in Japanese; Dautenhahn, 2007; Fong, Nourbakhsh, & Dautenhahn, 2003; Goodrich & Schultz, 2007). In 1970, Masahiro Mori, a Japanese robotics pioneer, proposed a hypothetical graph that predicted that the more human a robot looks, the more familiar it is to a human, until a tipping point is reached at which subtle nonhuman imperfections make the robot seem eerie. This 'dip' appears just before total human likeness (Mori, 1970/2012; MacDorman & Ishiguro, 2006). Mori cites dead bodies as an example of something that inhabits the uncanny valley, and he proposes that the eerie feeling associated with human-looking robots concerns the human instinct for self-preservation. For robot designers and computer animators, the uncanny valley poses an inevitable challenge to be overcome.

The uncanny valley has been examined from an evolutionary perspective (MacDorman & Ishiguro, 2006; MacDorman, Green, Ho, & Koch, 2009). Drawing on Rozin's theory, Keysers proposed the phenomenon could be associated with disgust, an evolved cognitive mechanism for pathogen avoidance (Curtis, Aunger, & Rabie, 2004; MacDorman & Ishiguro, 2006). We are more likely to be infected by the harmful bacteria, viruses, and other parasites of species that are closely related to us genetically; hence, we are most sensitive to signs of disease in our own species and least sensitive to signs of disease in animals that are only distantly related (Curtis, Aunger, & Rabie, 2004). Others have also proposed a relation between the uncanny valley and evolutionary aesthetics (MacDorman & Ishiguro, 2006). Our ancestors were under selective pressure to mix their

genes with the genes of those who could maximize the number and fitness of their progeny (MacDorman, Green, Ho, & Koch, 2009; Soler et al., 2003). The selective advantage of perceptual sensitivity to indicators of low fertility or a weak immune system could be responsible for the evolution of mechanisms underlying feelings of eeriness toward human forms that are sufficiently far from biological ideals (Rhodes & Zebrowitz, 2001).

2.2 Plotting Emotional Responses to Humanlike Characters

Assuming the uncanny valley proposed by Mori (1970/2012) exists (Figure 1), what dependent variables would be appropriate to represent Mori's graph? Mori referred to the dependent axis as *shinwakan,* a neologism even in Japanese, which has been variously translated as familiarity, rapport, and comfort level. Translating *shinwakan* as familiarity forges the link to Jentsch (1906) and Freud's (1919) seminal essays on the uncanny because in German, the language in which these essays were written, the uncanny *(das Unheimlich)* is constructed grammatically as an antonym of familiar *(das Heimlich)*. However, translating *shinwakan* as familiarity is problematic because familiarity cannot be equated with rapport or comfort level, and negative familiarity is undefined (Bartneck, Kanda, Ishiguro, & Hagita, 2009), given that zero familiarity is already total novelty (MacDorman & Ishiguro, 2006). Mori (1970) refers to negative *shinwakan* as *bukimi,* which translates as eeriness or uncanniness. In addition, prior to Ho and MacDorman (2010), no empirical scales have been developed to measure *shinwakan* in humanoid robots or other humanlike characters such as CG characters. Detailed questionnaires corresponding to the proposed benchmarks have not been developed and

tested empirically to show their reliability. However, the Godspeed questionnaire, compiled by Bartneck, Kulić, Croft, and Zoghbi (2009), includes five main concepts in human–robot interaction: *anthropomorphism, animacy, likeability, perceived intelligence,* and *perceived safety.* Although these researchers developed detailed semantic differential items for each concept, these indices have not been empirically tested for overall reliability and validity. In the study of Ho and MacDorman (2010), these indices are evaluated and then used to benchmark progress in developing a new set of indices.

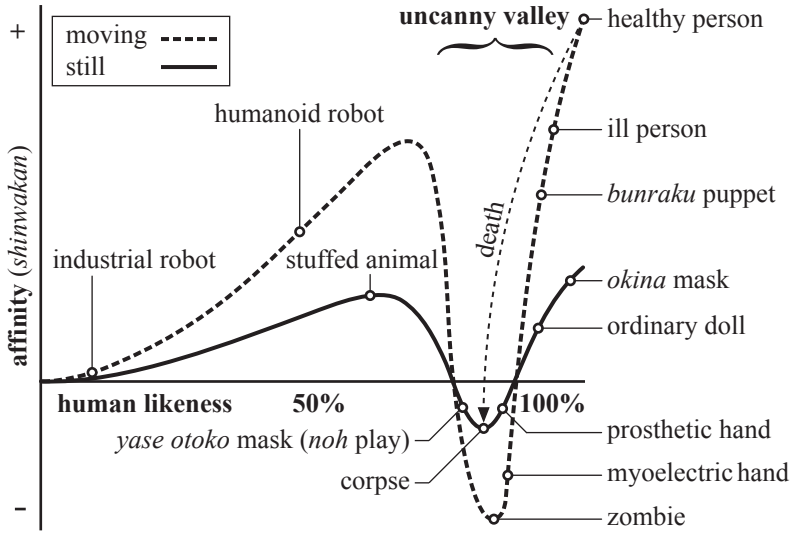

Figure 1. The uncanny valley (Mori, 1970/2012)

2.3 Studies on Emotion Similarity

Emotion researchers have tried to establish an emotion similarity space to see how we think about emotions based on empirical studies. They used statistical techniques to plot large sets of similarity judgments. The circular structure of the circumplex figure places emotions that are rated more similar closer together and emotions that are rated less similar farther apart (Larsen & Diener, 1992; Russell, 1980). In Figure 2, the first

dimension is *arousal:* Emotions involving high arousal can be grouped on one side of the circumplex, while those involving low arousal can be grouped on the other side. The second dimension is *valence*, which is orthogonal to arousal: Positive emotions are placed on one side, and negative emotions are place on the other side. The contribution of categorization showed a core relational theme associated with these emotions. For example, aroused represented excited, astonished represented surprised, and calm represented peaceful. They were unified by the fact that they are all positive emotions. The circumplex derived by Russell assumes these four conditions: (1) all items were extracted from just two dimensions; (2) items in each dimension have equal communalities; (3) all items are equally distributed in the space of the two dimensions; (4) any pair of two dimensions going through the space has equal distances (Acton & Revelle, 2000; Russell & Carroll, 1999).

Although Russell and his colleagues argue that inappropriate measurement masks the true bipolar structure of affect, providing additional support based on follow-up studies on different populations and cultures (Russell & Ridgeway, 1983; Russell, Lewicka, & Niit, 1989), the idea of bipolarity based on psychometric analysis is still being challenged (e.g., in neurology, psychopathology, and semantics; Cacioppo & Brentson, 1994; Rafaeli & Revelle, 2006; Watson, Wiese, Vaidya & Tellegen, 1999). Plutchik (1984) argued that all emotions could vary in arousal or intensity. For example, happiness can span from ecstasy to contentment, and anger can span from minor irritation to violent rage. Watson and Tellegen (1985) argued that positive and negative valences are independent instead of two ends of a common continuum. They reanalyzed some

early studies of self-reported moods by factor analysis to show positive and negative affects emerge as the first two dimensions with Varimax rotation method.

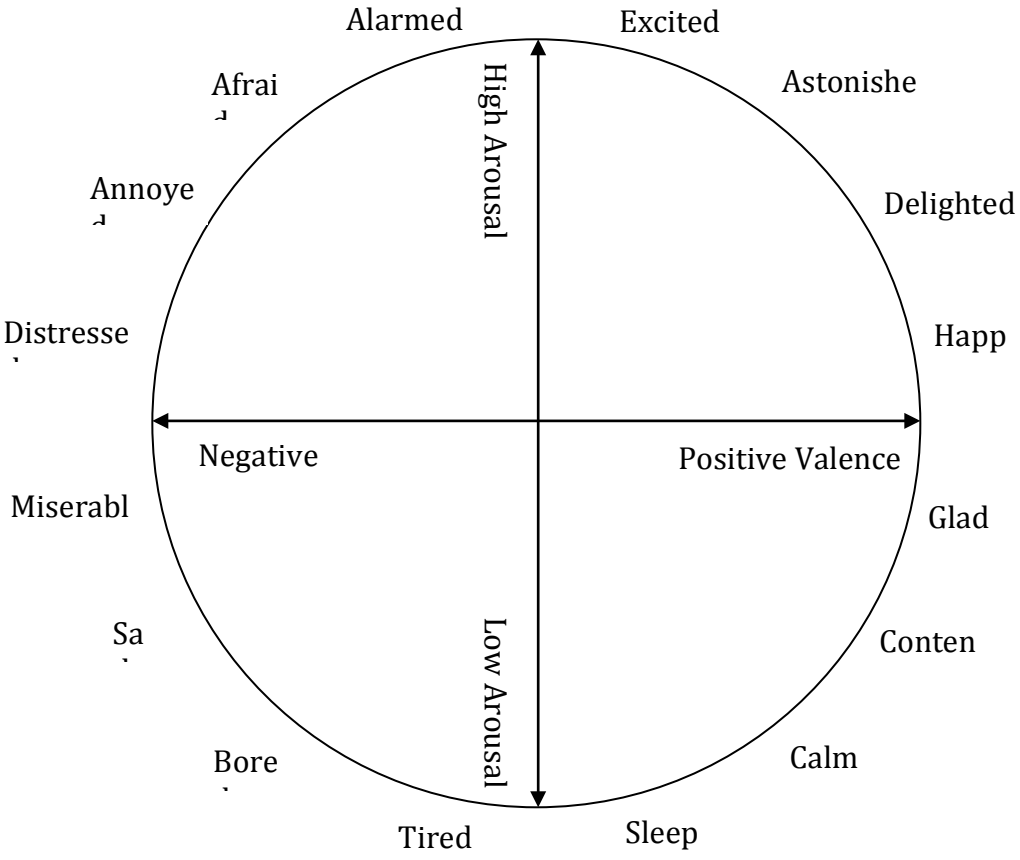

Figure 1: The circumplex of emotions, remade from Russell (1980)

Altarriba and Bauer (2004) used a comparison between emotion, abstract, and concrete words to examine the distinctiveness of emotion concepts. One of their interesting results showed that emotion words and abstract words mainly associated with words belong to the same type. In the word association experiment, participants could more easily recall emotion-related words in later recall as compared with other types of words. This result revealed that participants had greater agreements in emotional words than abstract and concrete words. It provides support for the use of emotion terms as valid instruments in this study.

Lane, Chua, and Dolan (1999) used positron emission tomography (PET) to measure regional cerebral blood flow while participants viewed neutral, pleasant, and unpleasant pictures, particularly in negative ones, which caused activations in the bilateral occipito-temporal cortex, left para-hippocampus gyrus, left amygdale, and cerebellum. Besides, Paradiso et al. (1999) found that pleasant pictures would cause more activity in neocortical areas than unpleasant pictures. Other studies showed that both negative and positive emotions caused neocortical activations. Northoff et al. (2000) used combined functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) techniques to determine that negative pictures would cause medial orbitofrontal activations and positive pictures would cause lateral oribitofrontal activation. However, these results were diverse. Some studies of negative emotions have found distinctive patterns of activation—but within the same anatomical region. For example, Philips et al. (1997) found with fMRI that perceiving a facial expression of disgust caused anterior insula activation. Liotti et al. (2000) found with PET memories of sad events caused activations in the right posterior insula, and memories of anxious events caused activations in the right ventral insula.

Prinz (2004) argued all emotions are compound. Some emotions may be intrinsically negative such as sadness or fear; some may be intrinsically positive such as joy or ecstasy, and some may have variable valance markers, such as surprise or curiosity. In some situations, both a negative emotion and a positive emotion were experienced concurrently. Some emotions were influenced by the recollection of past events, such as a mixture of joy and sadness while reminiscing on the past. Some mixed emotions are more dramatic. For instance, people may joyfully cry when reunited with long-lost relatives or

when winning the lottery. In addition, Provine (2000) found that jokes in our daily life cause only 20 percent of laughter; most of the times we laugh after hearing someone say something innocuous. Laugher is much like a social signal, which is constrained by social norms. Therefore, laughter and other expressions of happiness may not represent the original expressed emotion. This empirical evidence shows that emotions are highly mixed and associated with physical interactions and social circumstances.[4]

2.4 Positive-Negative Affect

Social psychologists have consistently found *warmth* and *competence* to be the two universal dimensions of human social cognition, when considering initial social perception of positive and negative affect (Fiske et al., 2007). Using a series of semantic differentials denoting traits, Asch (1946) found "striking and consistent" differences between affects when using "warm" versus "cold" as descriptive terms when social psychologists study how personality impressions are formed—as well as "competent" vs "incompetent." The two dimensions discovered by Asch using semantic differentials became a starting point for many other researchers, eventually taking the forms we know today as *warmth (vs. cold)* and *competence (vs. incompetence)*. Rosenberg and colleagues (1968) built on the work of Asch and also found two primary dimensions when forming impressions that they called "social" and "intellectual." The "social" dimension shares many traits with *warmth* and, in fact, socially desirable words clustered around "warm" and socially undesirable words clustered around "cold" when those words were plotted on an axis. Though both of these dimensions emerge at the time of first exposure,

[4] The section was credited to Ho, C.-C. (2008). Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings (Master's thesis, Indiana University). Retrieved from https://scholarworks.iupui.edu/

previous studies have shown that judgments of *warmth* are primary, emerging first and carrying more weight in affective and behavioral reactions (Fiske et al., 2007). Wojciszke and colleagues (1998) found *warmth* and *competency* together account for 82% of the variance in perceptions of everyday social behaviors.

Though *warmth* would appear to be a strong candidate for a measurement of immediate affect when first exposed, the original positive-and-negative measurement was not designed as semantic differentials. Recently, the *warmth* measure has been converted to semantic differentials and shows high internal reliability (Ho & MacDorman, 2010; Mitchell, Ho, Patel, & MacDorman, 2011). Using the bipolar semantic differential format to construct the alternative indices can reduce acquiescence bias without lowering psychometric quality (Friborg, Martinussen, & Rosenvinge, 2006; Lorr, & Wunderlich, 1988). Converting a Likert scale for warmth to a semantic differential scale allows comparison of the *warmth* measure to other well-known indices used to measure emotion such as the *pleasure, arousal,* and *dominance* (PAD) indices of Mehrabian and Russell (1974). Semantic differentials associated with *warmth* also tend to be strong measures of closely related kinds of positive affect such as *affinity, likability, communality, sociability, and comfortability* (Fiske et al., 2007; Abele & Wojciszke, 2007; Sproul et al., 1996; Wojciszke et al., 2009). As *warmth* encompasses such a broad range of measures, it is not surprising that indices for measuring the independent and dependent axes of Mori's graph all tend to be highly correlated with *warmth* and with each other. If indices are highly correlated with *warmth*, the positive and negative affect included in these indices might dilute their discriminative validity. Such correlation may also affect the orthogonal nature of the indices, making it difficult to plot the dependent variable(s) against the

independent variable(s), as Mori suggested (Ho & MacDorman, 2010). Therefore, in the development of psychological measurement, an inappropriate factor analysis might provide an improper model. Two such independent factors can be presented as two halves of one bipolar dimension, and vice versa.

2.5 Development of Humanness, Eeriness, and Attractiveness Indices

The Godspeed questionnaire, proposed by Bartneck, Kulić, Croft, and Zoghbi (2009), includes five main concepts in human–robot interaction: *anthropomorphism, animacy, likeability, perceived intelligence,* and *perceived safety*. They are not appropriate to represent their intended concepts because they are strongly correlated with each other and with positive and negative affect (Ho & MacDorman, 2010). The reason to reduce the influence of positive and negative affect in these indices is to be able to use each index independently. In other words, these indices should have the potential to be the standard benchmark for evaluating anthropomorphic entities. However, opposing semantic differential anchors should be designed to have roughly the same valence. This fact can be a challenge because human-related anchors tend to have a more positive valence than nonhuman or machine-related anchors. Based on previous studies of the uncanny valley, *humanness* can be an independent construct representing self-awareness, human awareness, and autonomy of anthropomorphic characters (Steinfeld et al., 2006); *eeriness* and *attractiveness* can measure the eeriness and comfort constructs proposed by Mori (1970/2012).

Ho and MacDorman (2010) used confirmatory factor analysis to test the Godspeed indices in the evaluation of various anthropomorphic characters. The results of

the validity analysis identified several other problems with the Godspeed indices: (1) the reliability of *Perceived Safety* was below the standard .70 cutoff; (2) the inconsistency of these indices demand several items be removed from the indices; (3) *Animacy*, *Likeability*, and *Perceived Intelligence* were deemed redundant owing to the high correlation among these indices. The high correlation indicates the Godspeed indices are measuring the same concept instead of their own presented concepts (*r* varied from .67 to .89 in the intercorrelations among *Anthropomorphism*, *Animacy*, *Likeability*, and *Perceived Intelligence*).

The multidimensionality of indices underlying the semantic differential technique is capable of demonstrating the perception of anthropomorphic characters (Gärling, 1976; Rosenberg et al., 1968). Indeed, in Ho and MacDorman (2010), *attractiveness, eeriness,* and *humanness* have high internal reliability in measuring anthropomorphic characters. These indices demonstrate a successful application of a bipolar semantic space to assess the perceived eeriness and comfort of anthropomorphic characters (Bentler, 1969; Lorr & Wunderlich, 1988; Russell, 1979). Several strengths of the empirical indices are shown: First, the *humanness* index that covers self-awareness, human awareness, and autonomy can measure human likeness based on appearance. It can also measure human likeness in the psychological sense (MacDorman & Cowley, 2006; MacDorman & Kahn, 2007). Second, both perceptual and emotional eeriness are relevant. Though correlated, they represent different concepts (Ho & MacDorman, 2008). In general, these indices are valid instruments for measuring their putative concepts in anthropomorphic characters.

Even though several advantages are present in the previously developed uncanny valley indices (Ho & MacDorman, 2010), the indices failed to distinguish between

humanlike robots and animated human characters. The scatterplots of *eeriness*, *humanness*, and *warmth* showed two clusters representing two kinds of human-looking entities. This is strong evidence against these two categories of anthropomorphic characters settling neatly into a continuum of human likeness. At certain boundaries (e.g., robot vs. animation, robot vs. human, or animation vs. human), incremental changes to human likeness in appearance (presentation) may produce disproportionately large changes in perceived category belonging. The ambiguous characters may lie on the cognitive boundaries. These cognitive boundaries may identify the feeling of the uncanny valley.

2.6 Cognitive Dissonance

Cognitive dissonance is the uncomfortable feeling that comes from holding two conflicting ideas. The ideas could be elicited by an anthropomorphic stimulus that lies on a category boundary—is it human or nonhuman, living or inanimate? Humans facing an unexpected stimulus change either their beliefs or behaviors to eliminate the inconsistency (Festinger, 1957). The negative feelings of cognitive dissonance are produced by competing alternatives regarding the categorization of the unexpected stimulus (Gerard & Mathewson, 1966; Joule & Azdia, 2003). This negative arousal might be evidence for cognitive dissonance as a cause of the uncanny valley effect (MacDorman et al., 2008; MacDorman et al., 2009). Human-looking interfaces could undermine current conceptions of personal and human identity. As human-looking computer interfaces become more humanlike, people may be challenged to see themselves more like machines. However, people may have difficulty in categorizing the

concept of the human-looking entity. Cognitive dissonance will result from the mismatch between a perceived human-looking computer interface and the already learned categories for people or other kinds of machines. Ramey (2005, 2006) stated that the uncanny valley is caused when two incongruent categories are joined by a quantitative metric that enables changes from one category to the other (e.g., human and robot). For instance, an observer may perceive a humanlike entity in the uncanny valley as familiar but strange. Unable to perceive whether the entity belongs to *human* or *robot*, the observer may eventually learn to identify the entity by a third category. However, this assumption has not yet been fully tested on adults and has only been measured using moderately humanlike robots (Kahn, et al., 2011, 2012; Kahn, Gary, & Shen, 2013). Other studies (Plantec, 2007, 2008; Tinwell & Grimshaw, 2009) attempted to use the paradox to explain the uncanny valley concerning CG characters in films and games. A human face sets up expectations about the associated voice and vice-versa; and the same applies for a robot face. In addition, a web experiment that created a mismatch in the audio and visual stimuli might also cause cognitive dissonance, which might explain why a mismatch has been found to increase eeriness and decrease warmth (Mitchell, Szerszen, Lu, Schermerhorn, Scheutz, & MacDorman, 2011).

However, with respect to the uncanny valley, cognitive dissonance may not result in a rationalization that pulls conflicting beliefs into alignment, because the origin of the category conflict may be largely perceptual and preconscious. Thus, cognitive dissonance might instead lead to an outright rejection of the object. In addition, an objective perception of the entity might decrease attributions of secondary emotions (e.g., admiration, resentment, love, or melancholy) to the outgroup member (Cortes et al.,

2005). The incongruity between human and nonhuman entities might trigger the objective perception of mind because nonhuman entities are perceived as being less mentally capable (Waytz, Gray, Epley, & Wegner, 2010).

2.7 Categorization Theories

Categorization is a basic cognitive ability that involves the comprehension of a different entity and a particular knowledge that includes both actual and potential instantiations (Croft & Cruse, 2004). The traditional view of conceptual categories is as fixed cognitive entities with stable associations with one or more linguistic expressions. However, recently emerging is the dynamic process of concept. It suggests that all aspects of conceptual categories are subject to revision. The prototype model of category structure (Rosch, 1973, 1978; Rosch & Mervis, 1975) attempts to measure two indicators: the goodness of exemplar (GOE) and the degree of membership (DOM). The GOE shows the frequency and order of mention, the order of learning, the structure of family resemblance, the speed of verification, and the magnitude of priming. The DOM includes three characteristics of concept: typicality or representativeness, closeness to an ideal, and stereotype. However, some problems of the simple prototype model might weaken its validity in practice. For example, it is insensitive to context. The relation between the number of features and GOE not only presents the availability of features but also reflects the presence of features dependent on the presence and the values of other features. In addition, once the participant develops a new category that covers the stimulus, the stimulus becomes self-evident, and the observer does not need to think about how to explain the stimulus by relating it to preexisting knowledge. At this point,

the participant's categorization process ceases. For example, when observers illustrate the categories of good and bad, they are limited in their descriptions of abstract categories. In addition, the contrasting category is another drawback for prototype theory. For example, observers may use *good* merely to show the opposite of *bad* without further explanation of the mutually exclusive relation between the terms (Croft & Cruse, 2004). The GOE and DOM indicate the difficulties inherent in measuring linguistic categories without a specific context.

2.8 Categorical Boundary

In cognitive science, *category* is an important concept to help an individual determine how to see and act. Some categories are innate—the result of evolutionary adaptation. For instance, infants can recognize different human faces (Ludemann & Nelson, 1988; Morton & Johnson,1991). In addition, some categories are determined by how culture and language subdivide concepts. For instance, certain color terms of various languages divide up color spaces differently and even within the same language, the usage may vary by social class (e.g., purple was an ecclesiastical color in the Medieval Age that was not generally worn by peasants). These inconsistencies indicate that categories are not only quantitative but also qualitative. Categorical perception occurs when the continuum of the perceptional dimension is judged as a series of discrete qualitative regions separated by the boundaries between labeled categories (Harnad, 1987). The distinguishability between animate and inanimate faces presents evidence for the existence of categorical boundaries along an anthropomorphism dimension (Looser & Wheatley, 2010). The divergent face pairs would increase the sensitivity of judgment that

passes over a tipping point. Human beings rely on facial cues for the categorical perception of animacy, especially in the area of the eyes. Furthermore, Kikutani, Roberson, and Hanley (2010) examined the learning effect of categorical perception. Their experiments indicate that human faces need to be categorized before the categorical perception can be established for the continuum between familiar and unfamiliar faces. Their experiments refute one assumption of the uncanny valley pertaining to the novelty of humanlike objects. Therefore, does the uncanny valley effect appear because of the ambiguity of two categories as seen in Figure 3 (e.g., a perceived android being perceived and understood with respect to *human* and *robot*)? In other words, does the uncanny valley effect merely happen to the individuals who cannot call up a new category label that can reduce the ambiguity (Ramey, 2006; Uekermann, Herrmann, Wentzel, & Landwehr, 2008)?

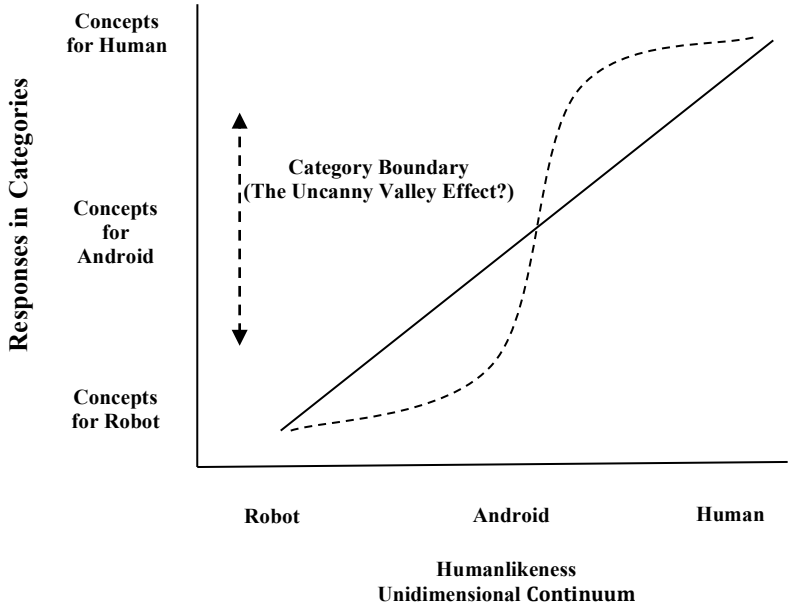

Figure 3. Categorical boundary in the uncanny valley

2.9 Card Sorting and Laddering Interview Techniques

Card sorting is a common method of usability testing to involve users in gathering information for a website (Capra, 2005; Dickstein & Mills, 2000; Rugg & McGeorge, 1997; Zimmerman & Akerelrea, 2002). Participants are asked to organize the content from the evaluated website in a way that makes sense to them. Participants review the items from the website and then group them into categories. Participants may even help label these groups. The strength of card sorting is the ability to build the structure for the website, decide the significant features put on the home page, and label the home page categories. The technique can ensure the organization of information on the website in a way that is logical to other users. Therefore, when applied to the evaluation of humanlike entities, this technique can reveal the underlying concepts of participant grouping of certain humanlike entities.

The laddering technique was originally used as a qualitative research technique to uncover the underlying reasons for people's behaviors. It refers to in-depth interviewing and analysis methods used to elicit the salient characteristics that customers seek when they make a choice to purchase a product (Reynolds & Gutman, 1988). More recently, the laddering method has been adapted for examining computer users' experience. It helps designers and researchers understand how well product attributes can facilitate personal values for end-users (Subramony, 2002; Zaman & Abeele, 2010). In the past decade, the focus in human–computer interaction (HCI) has extended from productivity to pleasure (Bødker, 2006). Therefore, the emphasis in usability studies has shifted toward users' subjective needs, referred to as user experience (UX). To analyze the usability and sociability of a product in different contexts, the laddering interview can be

combined with other techniques, such as the association technique, to understand users' needs comprehensively (Jans & Calvi, 2006). In addition, researchers can use a Web-based, visual technique for laddering to gather participants' responses instead of a locally administered technique. It quickly helps researchers understand the relation between interfaces and their users (Deutsch, Begolli, Lugmayr, & Tscheligi, 2011; Rugg, at el., 2002; Subramony, 2002). However, no studies have been performed in the area of human–robot interaction.

As foreseen, the laddering technique can reveal the underlying reasons why, for a given purpose, users interact with a particular robot over other alternatives. The advantages of the laddering method are its ability to show distinctions between subjects, to tell the order of priority, and to identify the importance in particular contexts (Reynolds and Gutman, 1988). Understanding the relations between a robot's attributes and humans' emotions can provide human–robot interaction researchers with useful information.

Laddering connects the values of users to their behaviors via a cognitive model of means–end chaining (Gutman, 1982). The central concepts of the means–end chain model are two linkages: the linkage of values and desired consequences and the linkage between consequences and product attributes. The means–end chain model is based on two fundamental assumptions about user behavior. First, users perceive and judge products as the "means" to achieve a desired "end-state" in a given product-use situation. Second, users cope with the overwhelming choices of products by grouping products into categories. For example, when the means–end chain model (attribute–consequence–value) is applied to the concept *robot*, users might think of categories labeled *humanoids*

and *androids*. However, they might also produce categories related to their functions and types of operation. Users' categorization may also include such groupings as *intelligent* or *automation*. The laddering interview technique can be used to examine the "ends," or values, in the means–end chain model that users believe when they interact with the robot. Laddering can also identify the categories in which users group the "ends." In marketing studies, laddering interviews consist of two steps: (1) eliciting salient characteristics and (2) probing to reveal the means–end structure (Hofstede et al., 1998). First, the participants identify attributes that distinguish different choice alternatives in a product class. This first phase is used to identify the available competitive set of products or services. For example, knowing which emotional responses or attributes that users use to infer the presence of desired consequences allows us to more clearly specify attribute development. Next, the laddering participants verbalize sequences of attributes, consequences, and values, which are referred to as ladders. Continuous probing is conducted by repeatedly asking a question such as "Why is that important to you?" (Reynolds & Gutman, 1988). This dialogue compels the participants to consider the reasons behind their choices or judgments—at least insofar as they are consciously accessible. These repetitive and probing questions reveal the means–end structure. For example, a robot designer could learn about specific emotional attributes that attract users to a robotic product or to the product of a competitor. These attributes can serve as indicators for the creation of meaningful association between the choice of a robot and the specific value that the user wants to gain. In the end, the responses of the individual ladder, or means–end chain, for each participant are aggregated and summarized.

2.10 Research Questions

Several research questions were addressed in this study: What is the relation between human categories and the perceptions of various humanlike forms? To what extent are these categories rooted in early “perceptual” or later “cognitive” processing? How do the categorical boundaries involve the emotional responses measured by the proposed indices?

3 METHODS

This study consisted of three phases to improve the attractiveness, humanness, and eeriness indices (Ho & MacDorman, 2010). The laddering interview was used to explore participants' concepts behind the categories of human-looking entities in Phase 1. The quantitative analysis of the laddering interview provided the item candidates of the indices in Phase 2. In Phase 3, a representative survey validated the indices based on the suggestions collected from the laddering interview.[5]

3.1 Participants

In Phase 1, 30 participants were recruited from a Midwestern university campus by email and flyers. Nine (30.0%) were female, 21 (70.0%) were male, and the median age was 26. Twelve (40.0%) were informatics majors and 18 (60.0%) were not. (Phase 2 is part of the analysis of Phase 1's data.) In Phase 3, the participants of web survey were recruited from an email list of randomly selected undergraduate students and recent graduates of a nine-campus Midwestern university system. Among the 1311 participants, 512 (39.1%) were male, 799 (60.9%) were female, 1068 (81.5%) were under 25 years old, 71 (5.4%) were 26–30, and 172 (13.1%) were over 31. The participants reflected the demographics of the university's undergraduate population. The measurement error range was ±2.89% at a 95% confidence level.

[5] The IUPUI/Clarian Research Compliance Administration approved this study (EX0903-35B). This experiment was supported by an IUPUI Signature Center grant. The laddering interviews were conducted from January 2013 to June 2013; the web survey was conducted from March 2014 to April 2014.

3.2 Materials and Procedures

In Phase 1, each participant viewed 12 video clips presented one at a time in random order. There were five video clips of three-dimensional computer animated characters, five of robots, and two of real humans (Figure 3). The video clips were 480 pixels by 360 pixels (a 4:3 aspect ratio). These clips were 15 to 30 seconds in length. After these clips were played, the participants were asked to categorize these 12 video clips and to group these clips by their categories. In the categorization task, the categories identified by the participant should be mutually exclusive. The participants were only allowed to assign one character into a unique category at a time. The categories cannot overlap. By presenting the pictures of these video clips as visual aids, participants were asked, "Which figures would you group together, or separate from others?" The participants were allowed to sort the figures into only one category. To increase the participants' recollection, they watched the clips based on their categories again. Then, participants completed a laddering interview on the figure featured in each video clip. Participants were asked repeatedly, "Why is that important to you?" Participants were required to provide at least three laddering responses. After the laddering interview, the participants rated on a 3-point scale (*not important, moderately important, very important*) all items of the attractiveness, humanness, and eeriness indices for each category the participant provided (Ho & MacDorman, 2010; Vanden Abeele, 1992).

In Phase 2, the participants' responses of laddering interview and item evaluation were converted into several data matrices for the analyses of hierarchical value map and new item candidate.

In Phase 3, the video clips and method of presentation were the same as in Phase 1. Each participant viewed 12 video clips presented one at a time in random order (Figure 4). Clips were played in a continuous loop while participants answered a survey on the figure featured in each clip. This round of the survey consisted of new items based on the candidates from Phase 2's results.

Figure 4. Twelve figures were rated by the participants: (1) Doctor Aki Ross and Captain Gray Edwards from the film *Final Fantasy: The Spirits Within;* (2) Billy, the baby from "Tin Toy"; (3) unnamed man from "Apology"; (4) Orville Redenbacher; (5) Mary Smith from "Heavy Rain: The Casting"; (6) iRobot Roomba 570; (7) JSK Laboratory's Kotaro; (8) Hanson Robotics' Jules; (9) David Ng's Animatronic Head; (10) Le Trung's Aiko; (11) Real Man; (12) Real Woman. No. 1 to 5 are animated figures; no. 6 to 10 are robotic figures; no. 11 and 12 are human figures.

3.3 Analysis

In Phase 1, to analyze the categorization procedure, participants' category responses were used to measure the network pattern of category boundaries, such as the number of categories, the size of each category, the heterogeneity of the categories, the

centrality of the category, and the dispersion of the category (Everett & Borgatti, 1999, 2005).

(1) The number of category: C

The number of categories was provided by the participant. For example, a participant might give two categories—human and robot—to group the figures. A larger number of categories meant that the participant had many categories for humanlike objects instead of just human vs. robot, or human vs. animation.

(2) The size of category: $\frac{F_i}{K_i}$

Where F_i was the number of figures that a participant categorized into a category, and K_i was the number of figures that were previously defined by the researcher (e.g., Figure no. 1 to no. 5 were computer-animated characters; Figure no. 6 to no. 10 were robots; Figure no. 11 and no. 12 were true humans). The size of each category indicated how likely the participants were to use these figures to represent the categories. If the size of the category was greater, it showed that the participant had broader categories elicited by the figure. For example, a participant identified five figures as belonging to the category of human, more than the two predefined human figures. This may indicate the participant's category of human was broader.

(3) Heterogeneity of categories: $\frac{C}{F}$

Where F was the number of figures and C was the number of categories given by the participant.

(4) Centrality of categories: $\frac{\sum S}{5C}$

Where C was the number of categories and S was the number of the figures correctly identified by a participant (e.g., the figure was a robot, and the participant identified it as a robot). Centrality indicates how close the category was. The answer ranges from 1 (very distance) to 5 (very close). If the centrality of category was larger, it showed that the participant associated the figures into the correct categories.

(5) Dispersion of categories: $1-\frac{1}{2N(N-1)/2}(\sum_{i=1}^{N-1}\sum_{j-1}^{N}S_{ij})$

The dispersion of the categories could be described as the relations in the whole category network. Dispersion represents the proportion of misidentified relations. To estimate this variable, the posited categories of the figures were used (e.g., Figure no. 1 to no. 5 were animated; Figure no. 6 to no. 10 were robotic; Figure no. 11 and no. 12 were truly human). Based on the correctness of the category task, it presented the dispersion of the categories. Where N was the number of figures within the category and S_{ij} was the similarity of category between figure i and j (Two figures were the same category = 1 and different category = 0). The range of dispersion was from 0 to 1.

In Phase 2, to analyze the laddering procedure, the taped interviews were transcribed. The first step in analyzing laddering data obtained from the research used a content analysis technique. Each idiosyncratic concept resulting from the laddering responses was categorized into one of three levels of abstraction—attributes, consequences, and values—in the means–end structure. Each element of the participants' responses should be included in one of these three categories. A number was assigned to each element to facilitate later coding.

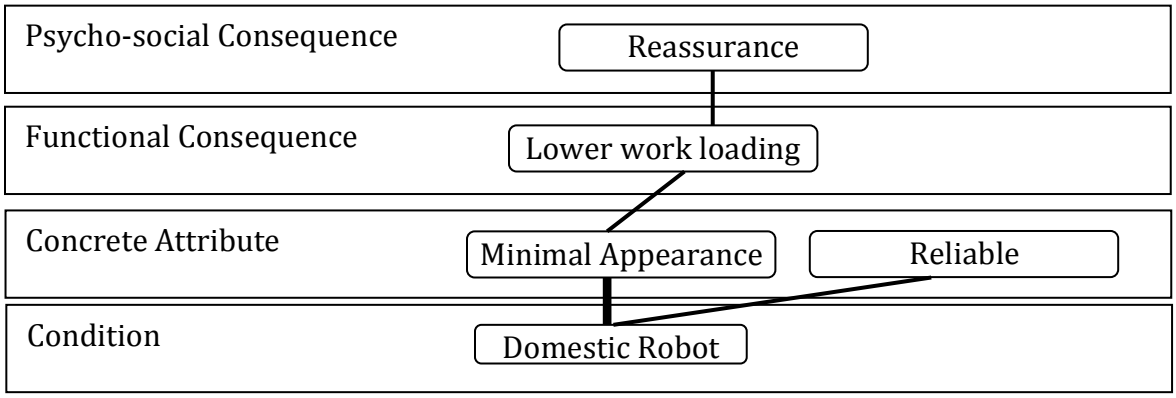

Figure 5. An example of a hierarchical value map

Therefore, a hierarchical value map (HVM) was constructed from the implication matrix as shown in Figure 5. An HVM was developed by connecting all the chains that were formed by considering the linkages in the large matrix of relations among elements. The HVM was able to show a well-organized summary of information derived in the interviews (Reynolds & Gutman, 1988). It provided a guide that showed what linkages of connecting values were important to the participant and to specific attributes of the product (Gutman, 1977). In addition, it presented the presence of desired consequences and attributes that permitted clearer concepts for the construction of a psychological index.

For the measurement purpose, the means-end chain generated a series of connected matrices from the HVM as values-by-consequences matrix, consequence-by-situations matrix, relevant consequences-by-grouping distinctions matrix, and relevant consequence-by-product matrix. The advantage of this approach was its ability to keep the illustration at a manageable size without becoming tangled in the methodology for generating the data (Gutman, 1982). These matrices were used to determine the distinctions comprising the chain and the connections between them.

To support index construction, after the laddering interview, participants evaluated the importance of each pair of attractiveness, eeriness, and humanness indices (Ho & MacDorman, 2010) as the candidates of the attribute (Claeys, Swinnen, & Vanden Abeele, 1995). The additional information helped the researcher to estimate the attributes gathered from laddering. Therefore, the attribute-consequence matrix (AC-matrix) and the consequence-value (CV-matrix) were conceived of as a series of connected matrices that could be used for index assessment (Hofstede et al., 1998). For the AC-matrix, the a priori attributes (the items of indices) and consequences were listed in the columns and rows, respectively, including all combinations of attributes and consequences. For the table of the CV-matrix, which included all possible combinations of consequences and values, the consequences and values were listed in the columns and rows, respectively. Although laddering was not intended to be used with representative samples, combining the item associations could uncover the concepts in AC- and CV-linkages.

To improve index construction, the linkages in AC- and CV-matrices from the laddering interview and the linkages in the importance of items were used to assess the convergent validity of the laddering interview and indices. With respect to convergent validity, the content of the laddering interview and the importance of item sorting were identified. The higher frequencies of item sorting also clarified the concepts gathered from the laddering interview.

From the laddering data, a three-way contingency table was generated and indexed by the attributes (A), consequences (C), and values (V). Hofstede et al. (1998) formulated a saturated model for the testing of the assumption. The probability P_{ijk} that a ladder consisted of attribute (i), consequence (j), and value (k) were expressed as a linear

equation in the parameters. The set of parameters consisted of a constant (α), the main effects (β_i^A, β_j^C, β_k^V), their interactions (γ_{ij}^{AC}, γ_{jk}^{CV}, γ_{ik}^{AV}), and the error (δ_{ijk}^{ACV}).

$$P_{ijk} = \alpha + \beta_i^A + \beta_j^C + \beta_k^V + \gamma_{ij}^{AC} + \gamma_{jk}^{CV} + \gamma_{ik}^{AV} + \delta_{ijk}^{ACV}$$

Based on the saturated model, the fitness of the laddering model was evaluated by means of the likelihood-ratio test statistic (χ^2). The corresponding likelihood-ratio test statistic was the chi-square difference test statistic $\Delta x^2 = x_{AC,CV}^2 - x_{AC,CV,AV}^2$. If the test was not significant, it meant the attributes and values were conditionally independent. It also supported the following analyses of AC- and CV-links in separate matrices.

Therefore, two saturated models were generated for testing the between-method convergent validity: the laddering interview and item evaluation. A new factor T was introduced for the measurement technique from which the laddering originates. The first was for the laddering interview; the second was for the item importance data.

$$P_{ijt}^{ACT} = \alpha + \beta_i^A + \beta_j^C + \beta_t^T + \gamma_{ij}^{AC} + \gamma_{it}^{AT} + \gamma_{jt}^{CT} + \delta_{ijt}^{ACT}$$

$$P_{jkt}^{CVT} = \alpha' + \beta'^C_j + \beta_k^V + \beta'^T_t + \gamma_{jk}^{CV} + \gamma'^{CT}_{jt} + \gamma_{kt}^{VT} + \delta_{jkt}^{CVT}$$

In these two models, β_t^T and β'^T_t represented the difference in the overall frequency of the concepts between two measurement techniques. The between-method difference in the frequency of occurrence of a specific attribute A_i, consequence C_j, or value V_k was taken into account by γ_{it}^{AT}, γ_{jt}^{CT}, γ'^{CT}_{jt}, and γ_{kt}^{VT}, respectively. The between-method difference in the frequency of A_iC_i-linkages was indicated by δ_{ijt}^{ACT}; the between-method difference in the frequency of C_jV_k-linkages was indicated by δ_{jkt}^{CVT}. Therefore, the terms, γ_{it}^{AT}, γ_{jt}^{CT}, γ'^{CT}_{jt}, and γ_{kt}^{VT}, represented the differences in the content of the cognitive network, whereas δ_{ijt}^{ACT} and δ_{jkt}^{CVT} reflected the difference in structure. By using these

saturated models, we were able to test the significance of the indices to validate the indices for the next phase.

In Phase 3, internal reliability was used to measure how reliable items were for their indices. Exploratory factor analysis, which applied principal components analysis with the Promax rotation, was used to verify that the semantic differential items loaded on factors corresponding to their named concepts. In addition, artificial–natural in the humanness index, reassuring–eerie in the eeriness index, and unattractive–attractive in the attractiveness index were chosen as "sanity check" items to verify the indices measured the concept after which they were named. Sanity check items had high face validity but did not necessarily meet the other criteria, such as being decorrelated with interpersonal warmth. If the results of factor analysis varied from the sanity check's dimension and showed low factor loadings, the items should be removed from the index. Correlation analysis would show the relation between indices and verify the discriminant validity of indices during testing. Confirmatory factor analysis would verify the theoretical structure of the new set of uncanny valley indices. Finally, multidimensional scaling was used to visualize similarities and dissimilarities among the semantic differential items by reducing the dimensionality of the space from higher dimensions to lower ones. Internal reliability, exploratory factor analysis, and correlation analysis was performed using SPSS, confirmatory factor analysis was performed using LISREL, and multidimensional scaling was performed using MATLAB.

4. RESULTS

4.1 Categorization

First, the summarized results revealed how the participants categorized the variety of anthropomorphic entities (Table 1). Although this study did not prevent the participants from categorizing all anthropomorphic characters into a single category, all of the participants proposed using at least two categories during the task. Of the 30 participants, more than half (54%) offered at least 4 categories (M=4.38) for the 12 characters, which is more than the three nominal categories of animations, robots, and humans. The categories mentioned most often were *Human, Robot*, *Animation*, *Machine*, *Woman*, *Man*, and *Android*. It showed that the participants would likely use more detailed categories to classify all of the anthropomorphic characters they saw. The participants were not satisfied with using broader terms, such as robot, for identification and wanted to use more specific terms like "advanced robot," "utility robot," and so on. Even though the participants had various identified categories, only a few used "humanlike robot" or "android" specifically.

Participants had quite different responses when identifying the specific category. For the animations, the participants only associated an average of 3.00 entities with the category of animation, which originally included 5 entities. Compared with the animations, the participants associated an average of 4.12 entities with the category of robot, which originally included 5 entities. Surprisingly, the participants associated an average of 2.88 entities with the category of human, which originally included 2 human entities. The results indicated that the participants would likely to group the animated characters into different categories (e.g., cartoon like, 3D computer-generated), a

consistent category of robots, and a broader category of humans. In other words, a realistic looking animated entity, or a sufficiently interactive robot, was likely to be categorized the same as human beings.

Considering the heterogeneity of categories, participants contributed an average of 7.33 categories for each entity. It indicated the participants had different thoughts on the entity they saw. Participants contributed an average of 6.2 categories for each animation entity, 9.6 categories for each robot entity, and only 4.5 categories for each human entity. This indicates the participants assigned many categories to the robots, perhaps because of their diversity in appearance or features, but used fewer categories on the humans. For the centrality of categories, the results were similar. Centrality was greatest for the human category (M = 4.58) and lower for the animation category (M = 3.16) and the robot category (M = 3.44). It indicated that the human category was the most robust. For the dispersion of categories, the average was .39. It indicated that the participants' categorizations were moderately loose identified relations. Unlike the categories of animation (M = .41) and robot (M = .49), the human category yielded only an average of .26 in category dispersion. This indicates the human category was unique from other categories.

Table 1. The most identified categories

Human (16)	Robot (15)	Animation (14)	Machine (5)
Woman (3)	Man (3)	Android (3)	Half human Half Robot (2)
Utility Robot (2)	3D Character (2)	Cartoon (2)	Advance Robot (2)
Prototype (2)	Humanlike Robot (2)	Machine Part (2)	Robot Machine (2)
Digital Creation (2)	Dummy (2)	Japanese Doll (2)	Advertisement (2)

Note. The value of each category represents the number mentioned by the participants.

Laddering responses were categorized into three kinds of comments: pro, neutral, and con, which was an efficient technique to measure their prevalence (Table 2). A pro comment meant that the participant's response valence was positive. A neutral comment meant that the participant's response was purely descriptive without valence. A con comment meant that the response's valence was negative.

For the animations, the average percentage of con was more than those of neutral and pro. Surprisingly, opinions of the animations became more bimodal as the interviews progressed. Although the percentage of pro comments was steady, the percentage of neutral decreased from 60.0% at the level of attribute to 27.5% at the level of value; the percentage of con increased from 37.5% at the level of attribute to 70.0% at the level of value.

For the robots, the comments kept steady from the level of attribute to the level of consequence. At the levels of attribute and consequence, the majority of comments were neutral. However, at the level of value, comments separated into pro, neutral, and con.

For the humans, the participants left only a small percentage of cons across the three levels. It indicated the participants viewed the category of human in a positive light. Especially in the value of human, 77.8% of comments were positive. It indicates that the participants inevitably linked the category of human with the concept of good. The results for the human category were the opposite of the animation category.

However, the participants viewed the androids differently from the others. At the level of attribute, the majority of comments about androids were pro (75.0%). Realistic appearance was described in the affirmative. However, at the levels of consequence and

value, the comments became more polarized. Half the comments about androids were positive, and the other half were negative.

Table 2. Pros, neutrals, and cons by different level of means-end

		Pro	Neutral	Con
Animation				
	Attribute	2.5%	60.0%	37.5%
	Consequence	2.5%	45.0%	52.5%
	Value	2.5%	27.5%	70.0%
Robot				
	Attribute	15.9%	58.7%	25.4%
	Consequence	12.7%	60.4%	27.0%
	Value	28.6%	36.5%	34.9%
Human				
	Attribute	38.9%	58.3%	2.8%
	Consequence	52.8%	41.7%	5.6%
	Value	77.8%	19.4%	2.8%
Android				
	Attribute	75.0%	12.5%	12.5%
	Consequence	12.5%	50.0%	37.5%
	Value	50.0%	12.5%	37.5%

4.2 Laddering Response

Several hierarchical value maps (HVMs) were constructed from the implication matrix that was coded by the laddering responses (Figure 6). The most frequent relations were used to illustrate the process by which the participants categorized animation, robot, and human. In this study, the cut-off value was 2 for 30 participants. If the value of each linkage was lower than 2, the relation was considered to be irrelative not to present in the

HVMs. For the category of animation character, three main attributes were "controlled," "computer generated," and "unreal." Three key consequences identified were "follow the plot," "no facial expression," and "unconvincing." Only two final values were contributed, "demonstration" and "soulless." Two complete means-end chains were found, "controlled–following the story–demonstration" and "computer generated–no facial expression–soulless." The incomplete means-end chain, "unreal–unconvincing," might be caused by the participant's inconstant responses in the level of value. It indicates the participants had the diverse ideas when observing that the animation character is unreal.

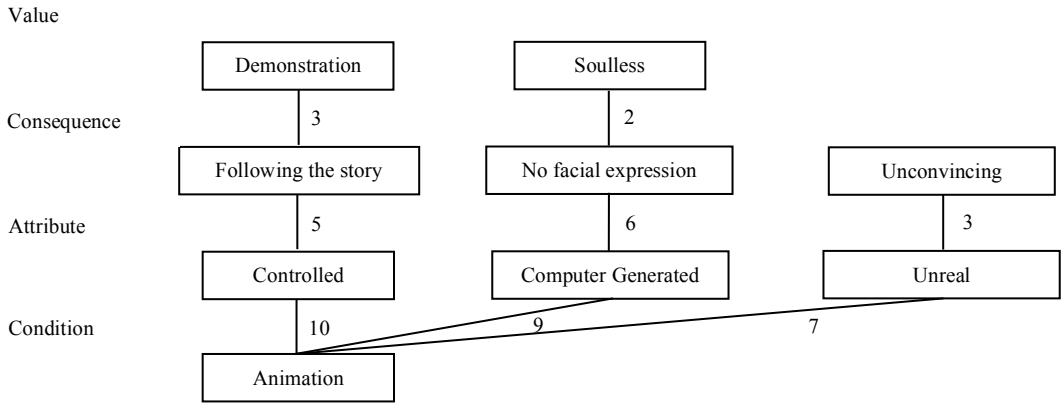

Figure 6. Hierarchical value maps of animation; the value of each linkage represents the number mentioned by the participants.

For the category of robot character, five key attributes were identified, "mechanical," "purpose served," "controlled," "interaction," and "human creation" (Figure 7). Five sequencing consequences were connected with their preceding attributes: "repeated movement," "doing its job," "technology," "convincing behavior," and

"machine." Although the participants identified many attributes and consequences, only three final values were linked, "demonstration," "simple work," and "no skin covering." The pattern indicated that the participants had stricter and more robust values toward the robots. This pattern was consistent with that of the categorization.

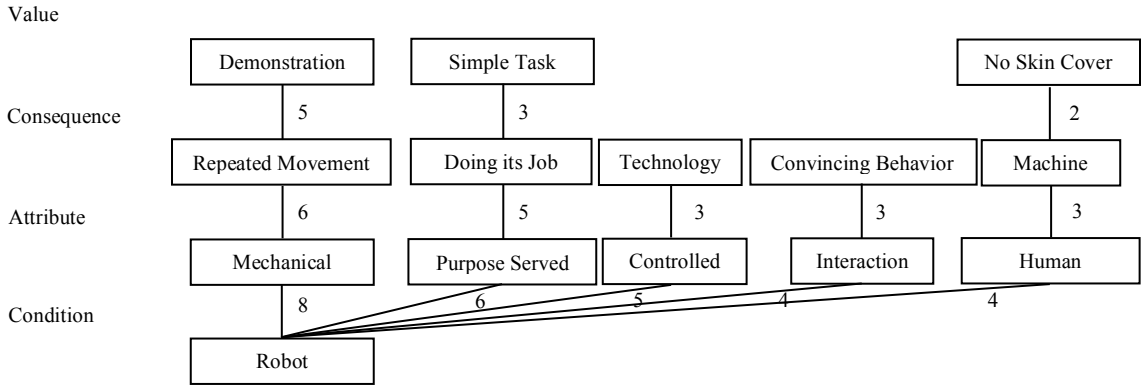

Figure 7. Hierarchical value maps of robot; the value of each linkage represents the number mentioned by the participants.

For the category of human character, three main attributes were identified by the participants, "interactive movement," "emotions," and "demonstration" (Figure 8). Three consequences were identified: "trust," "timing," and "convincing behavior." Only two final values were contributed in the end of laddering: "sophisticated" and "soul." Two complete means-end chains were found, "interactive movement–trust–sophisticated" and "demonstration–convincing behavior–soul." The results indicated the participants considered the humans had the qualities of interaction, trust, and sophistication.

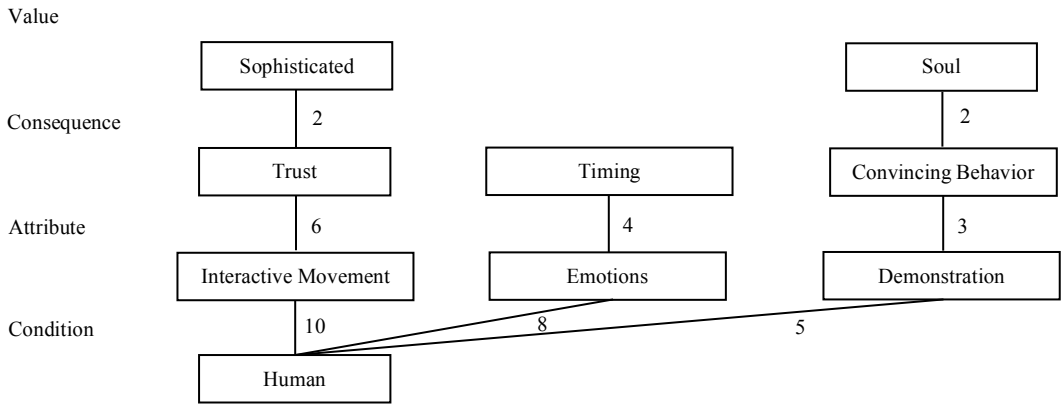

Figure 8. Hierarchical value maps of human; the value of each linkage represents the number mentioned by the participants.

In contrast, the android's hierarchical value map indicated mixed and conflicting thoughts (Figure 9). Three key attributes were identified, "convincing character," "humanlike appearance," and "unconvincing setting." Four consequences were connected with their preceding attributes, "convincing facial expression," "purpose served," "interaction behavior," and "post-production effect." Notably, the origin of "purpose served" and "interaction behavior" was "humanlike appearance." At the end of laddering, four final values were contributed: "emotions," "contingency," "mutual sense," and "personal experience." Four complete means-end chains were found in the matrix: "convincing character–convincing facial expression–emotion," "humanlike appearance–purpose served–contingency," "humanlike appearance–interaction behavior–mutual sense," and "unconvincing setting–post production effect–personal experience." However, these four means-end chains indicated that the participants had both positive and negative thoughts toward the androids. One positive chain indicated the android convinced the participants of its completed appearance and appropriate facial expression. The

participants were convinced that the android appealed to human emotions. Another negative chain indicated the participants suspected the android's imperfect setting, and they speculated upon any post-production effect. The participants imputed the unconvincing android to their personal experience. For example, the participants might have seen vicious androids in science fiction films and copied the idea to this category. A participant mentioned a vicious alien cyborg disguised as a female human in *Doctor Who* made him feel the android was scary.

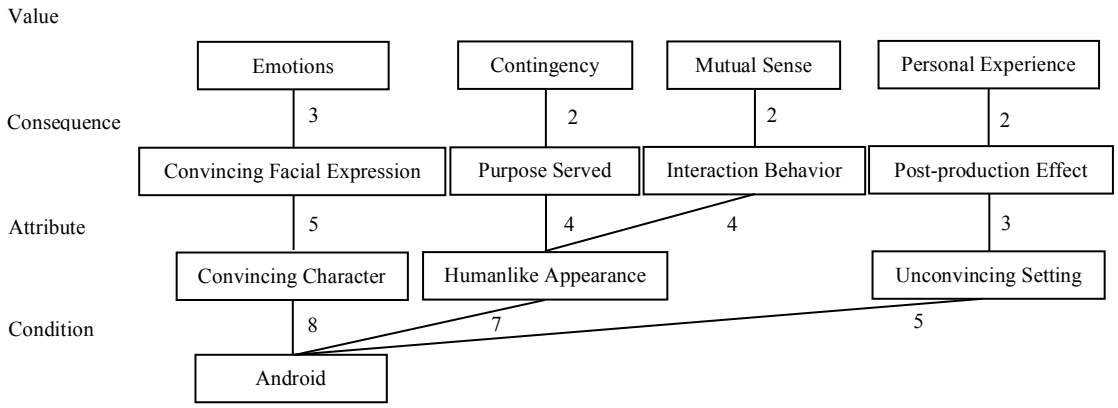

Figure 9. Hierarchical value maps of android; the value of each linkage represents the number mentioned by the participants.

4.3 Visualization of Categorical Boundaries

To illustrate the items forming categorical boundaries among the categories, a visual mapping of laddering interview were generated. A nodelist of 105 idiosyncratic items converted from the laddering responses was used to present the visual map (Figure 10).

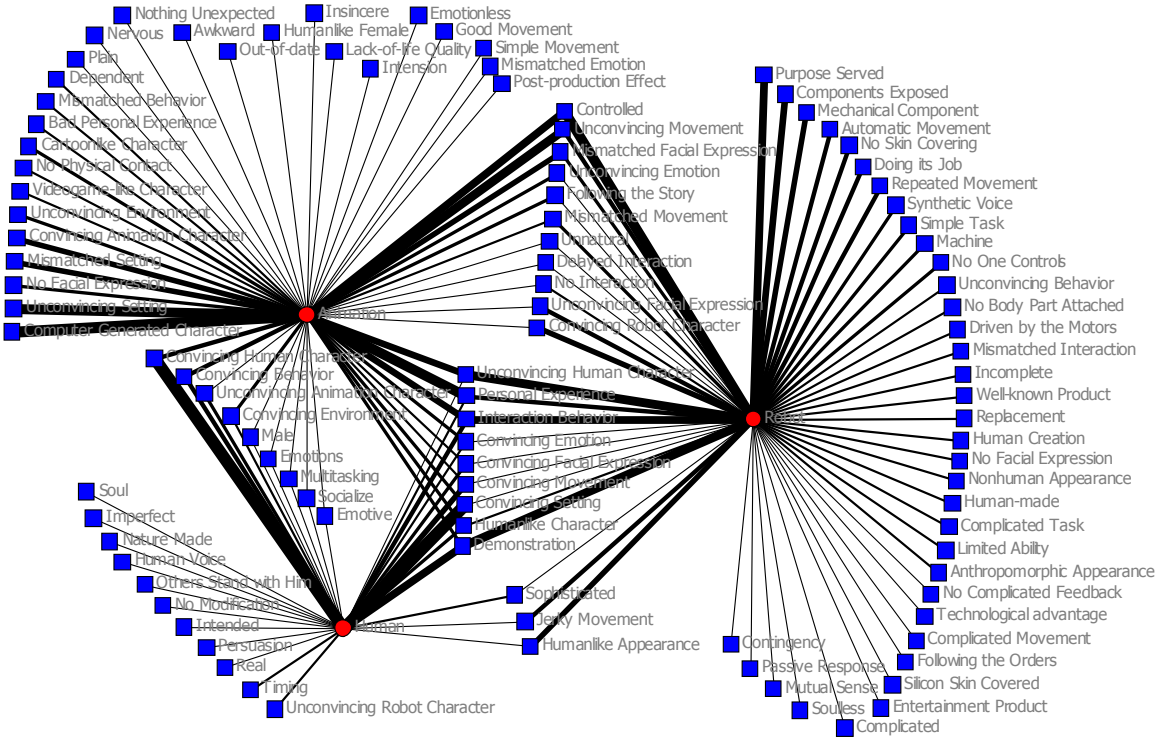

Figure 10. The visual mapping of idiosyncratic items associated with three posited categories

In the visual map, each coding item connected to three posited categories: the animation, the robot, and the human. The width of the line indicated the strength of association, operationalized as the co-occurrence frequencies of the associated items. Thicker lines indicated stronger associations; thinner lines indicated weaker associations. In general, 26 items were associated solely with the animation category (Figure 11); 36 items were associated solely with the robot category (Figure 12); and 11 items were associated solely with the human category (Figure 13). It indicates the animated characters and robots provoked more mental images than the human characters.

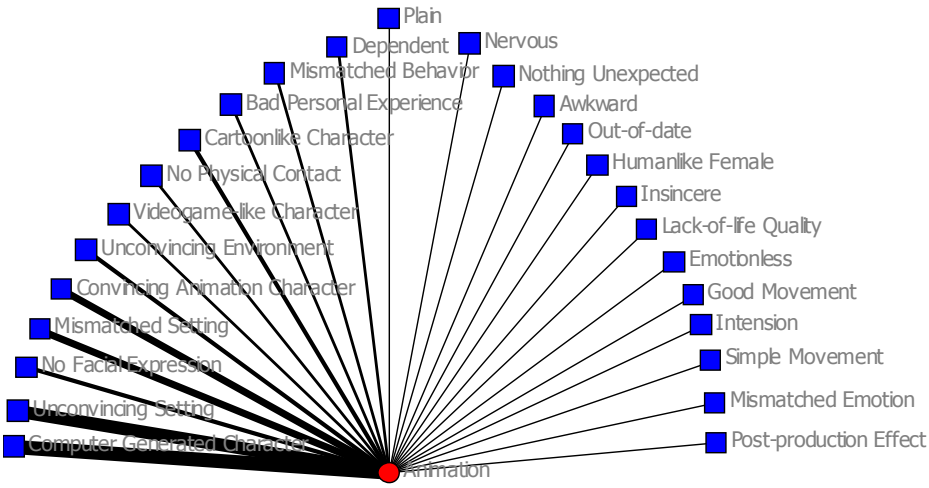

Figure 11. Idiosyncratic items solely associated with the posited animation category

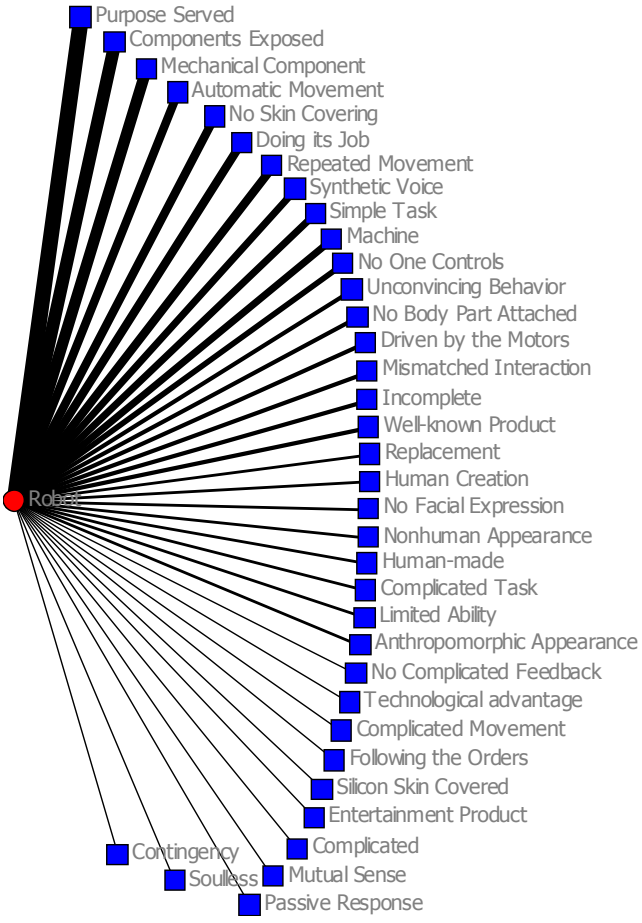

Figure 12. Idiosyncratic items solely associated with the posited robot category

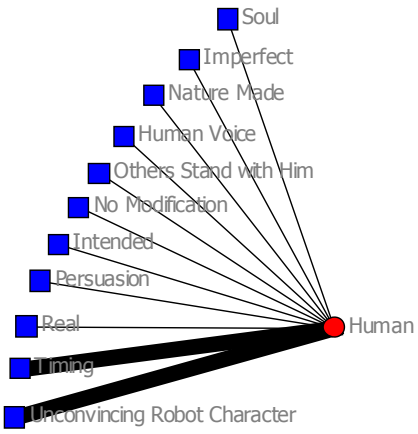

Figure 13. Idiosyncratic items solely associated with the posited human category

In the areas between the categories, three coding items linked the categories of human and robot: *Jerky Movement*, *Sophisticated*, and *Humanlike Appearance* (Figure 14). However, these three associations were asymmetric. *Jerky Movement* and *Humanlike Appearance* had stronger associations with the robot category but weaker associations with the human category. *Sophisticated* had a stronger association with the human category but weaker association with the robot category. This indicates a categorical boundary between human and robot. The participants occasionally associated *Humanlike Appearance* and *Jerky Movement* with robot but rarely with *Sophisticated*, and vice versa.

Figure 14. Idiosyncratic items coassociated with the posited human and robot categories

There were 9 coding items coassociated with the categories of human and animation: *Convincing Human Character*, *Convincing Behavior*, *Convincing*

Environment, *Unconvincing Animation Character*, *Emotions*, *Multitasking*, *Socialize*, *Male*, and *Emotive* (Figure 15). The associated pattern of *Convincing Human Character* and *Unconvincing Animation Character* were extremely asymmetric. They were strongly associated with the category of human but weakly associated with the category of animation. It indicated the participants were confident of the human category they identified; in the meantime, they denied every animation character's characteristic from the human category.

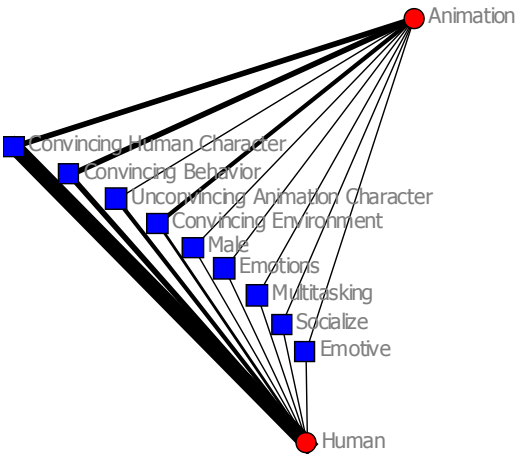

Figure 15. Idiosyncratic items coassociated with the posited human and animation categories

In addition, 11 coding items directly linked the categories of animation and robot together as *Controlled*, *Convincing Robot Character*, *Delayed Interaction*, *Following the Story*, *Mismatched Facial Expression*, *Mismatched Movement*, *No Interaction*, *Unconvincing Emotion*, *Unconvincing Facial Expression*, *Unconvincing Movement*, and *Unnatural* (Figure 16). In these associations, *Unconvincing Facial Expression* had stronger association with the robot category, but weaker association with the animation

category. It was opposite to *Mismatched Facial expression*. It indicated another categorical boundary about facial expression between robot and animated characters, χ^2 (1, N=20) = 5.50, p = .019. The participants likely considered the robots unable to perform appropriate facial expressions and the animation characters able to perform them but with the incorrect timing or action.

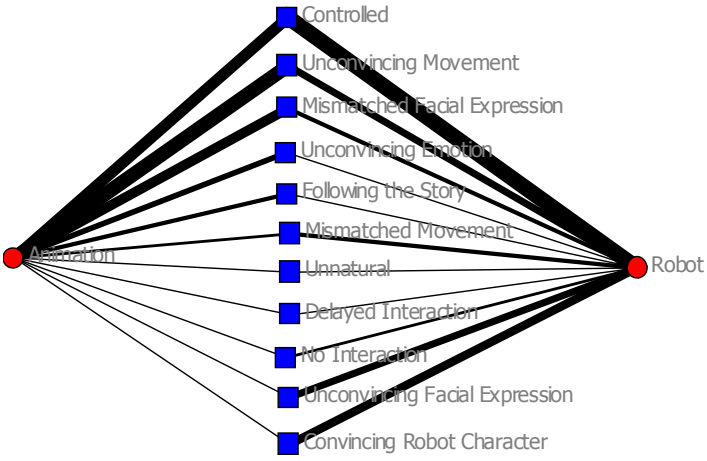

Figure 16. Idiosyncratic items coassociated with the posited animation and robot categories

Considering the self-identified categories, the alternative nodelist matrix was converted for the visualization. Four main self-identified categories, the animation, the android, the robot, and the human, were the roots connecting the coding items in this alternative visual map (Figure 17). In general, 17 items solely associated with the self-identified animation category (Figure 18); 30 items solely associated with self-identified robot category (Figure 19); 18 items solely associated with the self-identified human category (Figure 20); and 2 items solely associated with the self-identified android category.

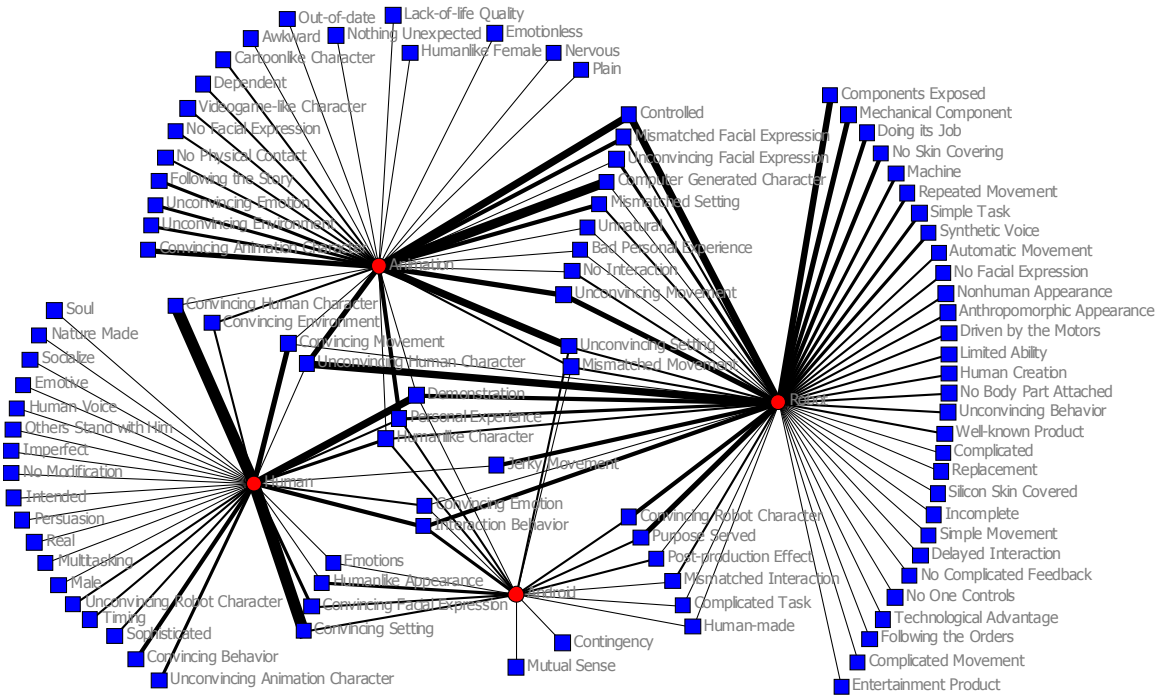

Figure 17. The visual mapping of idiosyncratic items associated with four self-identified categories

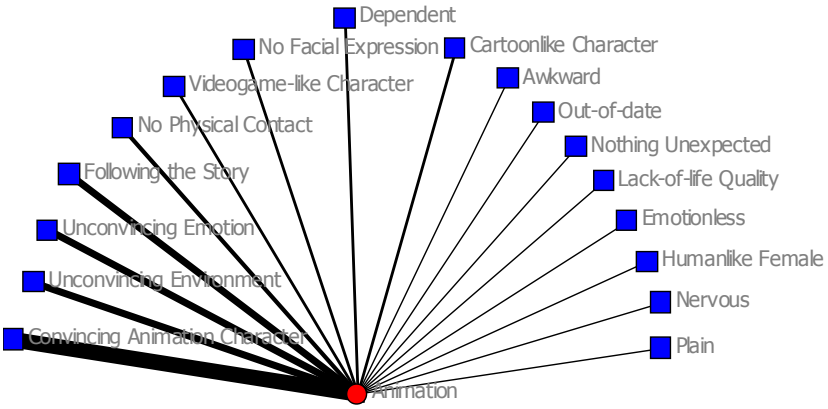

Figure 18. Idiosyncratic items solely associated with the self-identified animation category

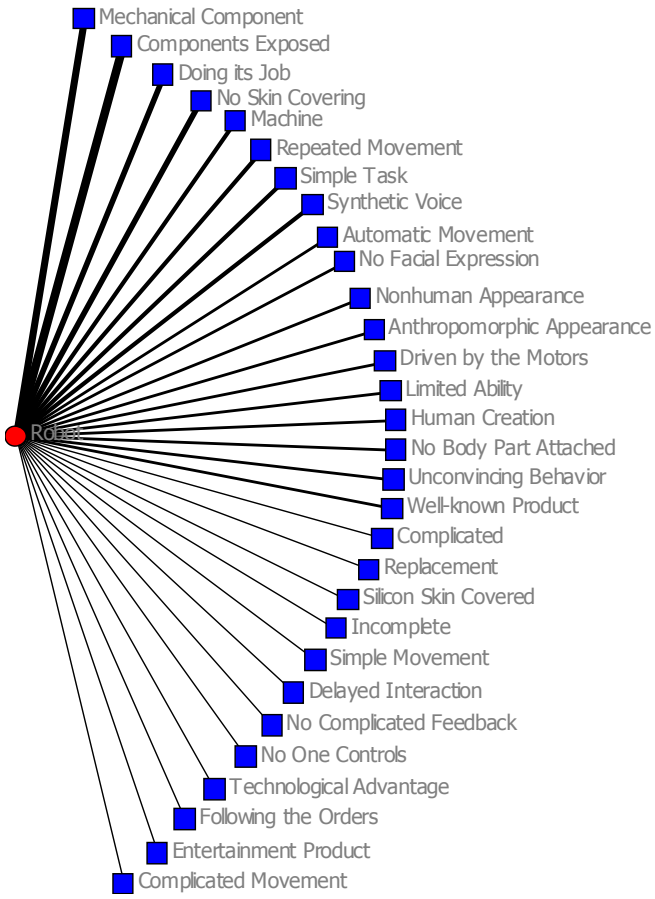

Figure 19. Idiosyncratic items solely associated with the self-identified robot category

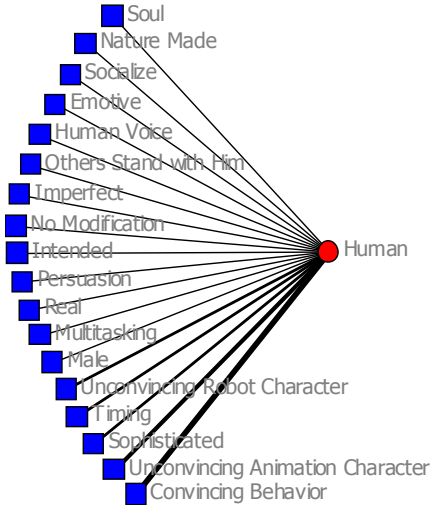

Figure 20. Idiosyncratic items solely associated with the self-identified human category

Compared with the mapping of the original category, the pattern of self-identified category seemed similar but different: *Demonstration, Personal Experience,* and *Humanlike Character* were commonly associated with all four categories (Figure 21). Unlike the mapping of the posited category, fewer associated items were linked with the categories. It indicated the participants would likely to use the particular idea on their self-identified categories instead of using the ambiguous one. *Jerky Movement* was the sole association between the categories of human and robot. Unlike the mapping of Figure 10, *Sophisticated* became one of the exclusively human characteristics. This indicates the participants used more unmixed characteristics for the human category. It also happened for the associations between the animation and human categories. Only *Convincing Human Character* and *Convincing Environment* became the direct associations instead of *Convincing Behavior*, *Unconvincing Animation Character*, *Male*, *Multitasking*, *Emotive*, and *Socialize*. This indicates the participants considered more exclusively human characteristics when they gave their own categories.

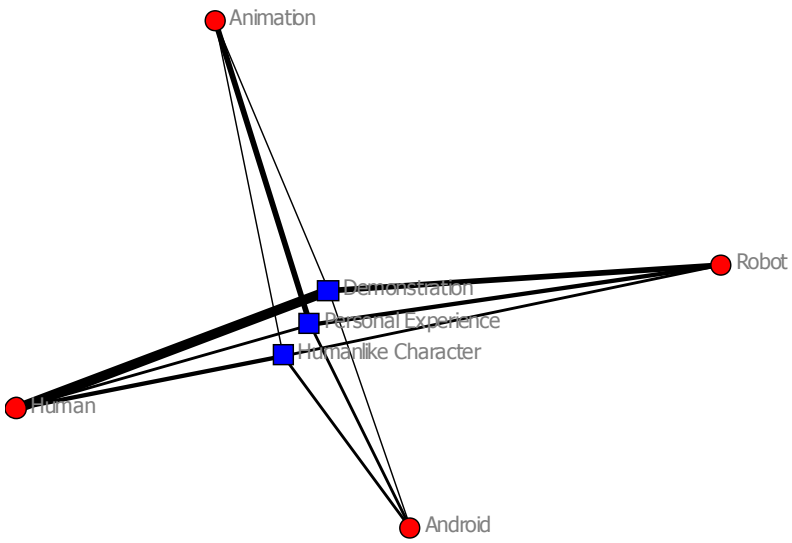

Figure 21. Idiosyncratic items coassociated with the four self-identified categories

For the new self-identified category of android, two items, *Contingency* and *Mutual Sense* emerged from the associations. They were solely associated with the android category. It revealed how the participants conceived the idea of android. The android was capable of detecting human's contingent responses and giving the mutual sense to the humans. In addition, four associations, *Emotions*, *Convincing Setting*, *Convincing Facial Expression*, and *Humanlike Appearance*, became the shared associations between the human and android categories (Figure 22). In the mapping of the original category, *Convincing Setting* and *Convincing Facial Expression* were the common items associated with three categories; *Emotions* was the association between the animation and human categories; *Humanlike Appearance* was the association between the human and robot categories. This indicates the android category, which the participants identified, borrowed most of its associations from the animation, robot, and human categories; decreasing the ambiguities in the context.

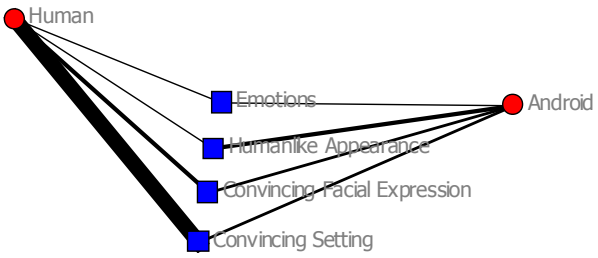

Figure 22. Idiosyncratic items coassociated with the self-identified human and android categories

4.4 Item Evaluation

After the card sorting, the evaluations of 38 terms based on the perceived categories were analyzed, which were decomposed by the 19 semantic differential scales (Ho & MacDorman, 2010). Among the 38 terms of the humanness, eeriness, and attractiveness indices, humanness items were the most important to all identified categories (M=2.0, SD=.25); attractiveness items were the second most important (M=1.64, SD=.40); and eeriness items were least important (M=1.60, SD=.33). The result showed that the participants were more sensitive to the items of the humanness index. When comparing the positive and negative terms in each index, the preference showed how the participants would be likely to use the items to categorize the anthropomorphic characters. The results indicate that when categorizing the anthropomorphic characters, the participants were more likely to choose fewer humanness (M=–0.34, SD=1.24), fewer eeriness (M=0.24, SD=0.63), and more attractiveness (M=0.33, SD=0.82) terms.

In addition, we compared the importance and preference of terms by different categories such as robot-related categories versus others, animation-related categories versus others, and human-related categories versus others (Table 3). In the importance of terms, the participants who identified animation-related categories (M=1.87, SE=.07) would likely use fewer humanness terms than those of other (M=2.03, SE=.03) categories (F(1, 61)=4.37, p=.041). The participants who identified human-related categories (M=1.85, SE=.12) would likely use more attractiveness terms than those of other (M=1.57, SE=.05) categories (F(1, 61)=6.18, p=.016). In the preference of items, the participants who identified robot-related categories (M=–1.12, SE=.16) would use fewer humanness terms than other (M=–0.11, SE=.20) categories (F(1, 61)=18.57, p<.001). The

participants who identified human-related categories (M=1.56, SE=.11) would use more humanness terms than those of other (M=–0.94, SE=.10) categories (F(1, 61)=172.93, p<.001). The participants who identified human-related categories would use fewer eeriness terms (F(1, 61)=12.47, p<.001). The participants who identified human-related categories would use more attractiveness terms (F(1, 61)=7.91, p=.007).

Table 3. The importance and preference by categories

	Importance					
	Humanness		Eeriness		Attractiveness	
	M	*p*	*M*	*p*	*M*	*p*
Robot-related	2.07	.105	1.70	.084	1.62	.788
Others	1.96		1.55		1.66	
Animation-related	1.87	.041	1.53	.306	1.52	.386
Others	2.03		1.62		1.67	
Human-related	2.09	.123	1.62	.802	1.85	.016
Others	1.97		1.60		1.57	
	Preference					
	Humanness		Eeriness		Attractiveness	
	M	*p*	*M*	*p*	*M*	*p*
Robot-related	–1.12	.000	–0.15	.388	0.22	.400
Others	0.11		–0.29		0.40	
Animation-related	–0.82	.117	–0.12	.436	0.15	.380
Others	–0.21		–0.27		0.38	
Human-related	1.56	.000	–0.70	.001	0.83	.007
Others	–0.93		–0.09		0.18	

4.5 Revised item suggestion

When the participants identified various categories, the participants' ratings of the indices would be influenced by the categories they selected. The participants would likely

consider the categories they identified instead of the items of the indices they evaluated. To reduce the bias, we compared each semantic differential pair by different categories. The results indicated that the pair "Without Definite Lifespan–Mortal" (p=.006) of the humanness index was significantly biased by the category of robot; the pair "Numbing–Freaky," (p=.005) and "Unemotional–Hair-raising" (p=.002) were significantly biased by the category of robot. For the category of animation, two pairs "Synthetic–Real," (p=.007) and "Mechanical Movement–Biological Movement" (p=.014) of the humanness index, were significantly biased. For the category of human, the pair of the humanness index "Inanimate–Living" (p=.001) was significantly biased. Three pairs of the eeriness index, "Reassuring–Eerie" (p=.007), "Ordinary–Supernatural" (p=.000), and "Unemotional–Hair-raising" (p=.019), were significantly biased. Two pairs of the attractiveness index, "Unattractive–Attractive" (p=.034) and "Crude–Stylish" (p=.013), were significantly biased. For the category of android, two pairs of the eeriness index, "Numbing–Freaky," (p=.014) and "Unemotional–Hair-raising" (p=.029) were significantly biased. The results suggested these terms needed further revising to eliminate the subjective bias. Considering the results of the participants' item evaluations, three pairs of the eeriness index, "Numbing–Freaky," "Ordinary–Supernatural," and "Unemotional–Hair-raising" would likely be biased across various categories.

Using the participants' laddering responses as the pool of item suggestion, "Numbing–Freaky" could be revised as "Dull–Freaky" or "Boring–Freaky"; "Ordinary–Supernatural" could be revised as "Ordinary–Unreal" or "Ordinary–Creepy"; "Unemotional–Hair-raising" could be revised as "Unemotional–Alarming"; "Reassuring–Eerie" could be revised as "Predictable–Eerie." In addition, "Plain–Weird," "Conformist–

Bizarre," and "Habitual–Supernatural," were also considered potential pairs based on the participants' laddering responses. These newly revised items will be applied in the following web survey with the original ones to test whether they are more appropriate.

4.6 Conditional Independence

From the laddering interview, the complete attribute-consequence-value linkages were recorded, Hundreds of laddering responses were coded: 150 ladders were in the posited animation category; 150 ladders were in the posited robot category; and 60 ladders were in the posited human category. While considering the participant's self-identified category, 104 ladders were in the self-identified animation category; 144 ladders were in the self-identified robot category; 91 ladders were in the self-identified human category; and 21 ladders were in the self-identified android category. The number of different attributes, consequences, and values after content coding were shown in Table 4. These ladders were used to construct a 3-way $A_iC_jV_k$ contingency table for each category. The cells for the 3-way table presented the frequencies with which each of the linkages occurred in the laddering data.

Table 4. Characteristics of the laddering data used to test conditional independence

		Interviews	Attributes	Consequences	Values
Posited	Animation (5)	150	22	22	34
	Robot (5)	150	29	31	36
	Human (2)	60	8	12	15
Self	Animation (Mode=4)	104	14	15	20
	Robot (Mode=5)	144	30	29	34
	Human (Mode=3)	91	10	11	20
	Android (Mode=3)	21	6	7	8

The normed fit indices Δ and the adjusted χ^2 statistics for the tests of conditional independence were indicated in Table 5. The adjusted χ^2 statistics for the posited categories and the self-identified models were low. These results indicated the models fit the data well. The normed fit indices were close to 1. This indicates little space for improvement. The *p*-values were insignificant. This indicates strong empirical evidence for the independence of both AC- and CV-matrices. It indicated that the attributes and values are associated indirectly through the attribute-consequence and consequence-value linkages. In addition, comparing these statistics of the original categories with those of the self-identified ones, the self-identified categories' Δs and the adjusted χ^2 would fit the model better. The results indicated the participant's prior categorization facilitated the validity of the laddering interview.

Table 5. Tests for conditional independence of attributes and values given the consequences

	AC, CV, AV			*AC, CV*			*AC, CV vs. AC, CV, AV*		
	Δ	x^2_{adj}	*df*	Δ	x^2_{adj}	*df*	Δx^2_{adj}	*df*	*P*
Posited									
Animation	.87	66.57	2850	.84	466.97	3542	288.44	692	>.999
Robot	.82	778.07	4498	.78	272.57	5199	540.46	701	>.999
Human	.89	28.77	204	.91	46.81	250	31.08	46	.954
Self-identified									
Animation	.83	54.94	1819	.81	374.8	2410	270.66	591	>.999
Robot	.91	264.58	3851	.90	278.06	4201	271.52	350	>.999
Human	.95	30.21	330	.90	44.88	385	36.03	55	.978
Android	.90	9.22	85	.89	25.54	112	16.01	27	.953

4.7 Between-Method Convergent Validity

Based on the previous test, the assumption of conditional independence was supported; the two models of P_{ijt}^{ACT} and P_{jkt}^{CVT} can be estimated separately. The evaluation of item importance was considered as the external technique to test between-method convergent validity. To test the convergent validity of the laddering interviews and evaluated indices, all AC- and CV-linkages from the laddering interviews were used to compare the item importance toward different categories with the number of direct and indirect AC- and CV-linkages from the laddering matrix. Both the laddering interview and the item evaluation were combined and transformed in two 3-way contingency tables

containing the frequencies of the A_iC_i- and C_jV_k-linkages respectively. Several tests for similarity in content were significant (p<.001), including the attributes ([A, C, T] vs. [C, AT]), consequence ([A, C, T] vs. [A, CT]), and attributes and consequence simultaneously ([A, C, T] vs. [AT, CT]). It indicated that the content of the AC-matrix significantly differed between the laddering interview and item evaluation (Table 6). In addition, in the test for structural similarity, the model ([AC, AT, CT]) vs. the saturated model ([A, C, T]) was insignificant. This indicates the laddering interview and items of indices might be similar in terms of structure. For the model test of the CV-linkage, the result was similar to the model test of the AC-linkage (Table 7). Based on the between-method convergent validity tests, both techniques were capable of exploring the concepts of various anthropomorphic categories.

Table 6. Tests for the similarity of laddering and item evaluation based on the AC-linkage

Model	Δ	x^2	*Df*	*p*	Test	Δx^2	*df*	*P*	Test for
[A, C, T]	.72	1219.81	380	<.001					
[C, AT]	.73	1156.32	371	<.001	[A, C, T] vs. [C, AT]	63.49	9	<.001	Content
[A, CT]	.73	1120.69	355	<.001	[A, C, T] vs. [A, CT]	99.12	25	<.001	Content
[AT, CT]	.76	989.54	349	<.001	[A, C, T] vs. [AT, CT]	230.27	31	<.001	Content
[AC, AT, CT]	.88	171.81	149	.097	[AC, AT, CT] vs. [A,C,T]	171.81	149	.097	Structure

Table 7. Tests for the similarity of laddering and item evaluation based on the CV-linkage

Model	Δ	x^2	*Df*	*p*	Test	Δx^2	*df*	*P*	Test for
[C, V, T]	.69	2681.95	684	<.001					
[V, CT]	.75	2029.26	662	<.001	[C, V, T] vs. [V, CT]	352.69	22	<.001	Content
[C, VT]	.71	2353.81	668	<.001	[C, V, T] vs. [C, VT]	328.14	16	<.001	Content
[CT, VT]	.78	1711.64	650	<.001	[C, V, T] vs. [CT, VT]	970.31	34	<.001	Content
[CV, CT, VT]	.89	361.06	334	.148	[CV, CT, VT] vs. [C,V,T]	361.06	334	.148	Structure

4.8 Validation of New Items

(1) Attractiveness Index

The four items of the attractiveness index were validated together: *Ugly–Beautiful*, *Repulsive–Agreeable*, *Crude–Stylish*, and *Messy–Sleek*, and the sanity check *Unattractive–Attractive*. The overall internal reliability of the index was high (Cronbach's α=.85). The exploratory factor analysis showed all four items including the sanity check loaded on a single factor that explained 65.08% of the variance. It confirmed the reliability of the original attractiveness index (Ho & MacDorman, 2010).

(2) Humanness Index

Similar to the attractiveness index, the five items of the humanness index were validated together: *Synthetic–Real, Inanimate–Living*, *Human made–Humanlike*, *Mechanical Movement–Biological*, and *Without Definite Lifespan–Mortal*, and the sanity check *Artificial–Natural*. The overall internal reliability was high (Cronbach's α=.84). The exploratory factor analysis showed all five items including the sanity check loaded on a single factor that explained 58.30% of the variance. It also confirmed the reliability of the original humanness index on similar samplers (Ho & MacDorman, 2010).

(3) Eeriness Index

First, all seven items of the original eeriness index and its sanity check were validated. Factor analysis confirmed the existence of the two subdimensions of the eeriness index previously found in Ho and MacDorman (2010). *Uninspiring–Spine-tingling*, *Boring–Shocking*, *Predictable–Thrilling*, *Bland–Uncanny*, and *Unemotional–Hair-raising* loaded on the first dimension, which explained 39.54% of the variance. The internal reliability of the first dimension was .84. *Reassuring–Eerie*, *Numbering–Freaky*,

and *Ordinary–Supernatural* loaded on the second dimension, which explained 23.62% of the total variance. However, the internal reliability of the second dimension was .69, indicating some space for improvement.

Considering the potential items of the eeriness index, nine new item candidates still followed the pattern of two dimensions as well as the original index. Seven item candidates, *Dull–Freaky*, *Ordinary–Unreal*, *Ordinary–Creepy*, *Plain–Weird*, *Predictable–Eerie*, *Conformist–Bizarre*, and *Habitual–Supernatural*, loaded with the dimension of *Reassuring–Eerie*, *Numbing–Freaky*, and *Ordinary–Supernatural*. Two item candidates, *Unemotional–Alarming* and *Boring–Freaky*, loaded with the dimension of *Uninspiring–Spine-tingling*, *Boring–Shocking*, *Predicable–Thrilling*, *Bland–Uncanny*, and *Unemotional–Hair-raising*.

First, the two candidates of *Ordinary–Creepy* (r=.70) and *Habitual–Supernatural* (r=.71) were highly correlated with the dimension of eerie, respectively. They were the redundant items or overlapped with other items. Therefore, *Ordinary–Creepy* and *Habitual–Supernatural* were excluded. Second, adding two candidates of *Unemotional–Alarming* and *Boring–Freaky* only increased the internal reliability of the dimension of spine-tingling (Cronbach αs ranged from .84 to .86). This indicates the dimension of spine-tingling, which included *Uninspiring–Spine-tingling*, *Boring–Shocking*, *Predicable–Thrilling*, *Bland–Uncanny*, and *Unemotional–Hair-raising*, had already saturated. Given that these five reliable items to measure the concept were already available, we did not need to develop additional items. *Unemotional–Alarming* and *Boring–Freaky* were excluded from the final index. Third, checking the correlations between the attractiveness and humanness indices, *Ordinary–Creepy* ($r_{\text{Attractiveness}}$=–.45 &

$r_{Humanness}$=–.31), *Ordinary–Unreal* ($r_{Attractiveness}$=–.37 & $r_{Humanness}$=–.44), *Conformist-Bizarre* ($r_{Attractiveness}$=–.35 & $r_{Humanness}$=–.28), and *Numbing-Freaky* ($r_{Attractiveness}$=–.30 & $r_{Humanness}$=–.23) were significantly correlated with the attractiveness and humanness indices, which violated the criterion of item decorrelation. Therefore, they were excluded from the final index.

Based on the three criteria of item selection (i.e., high internal reliability, correct factor loading, and correlation with the "sanity check" item), four items were constructed for the final version of the attractiveness index; nine items were constructed for the eeriness index; and five items were constructed for the humanness index. Confirmatory factor analysis was used to test the theoretical structure of the final set (Table 8). It showed the factor loadings for the 18 semantic differential items of the final set. Although one goodness-of-fit index (RMSEA = .061) slightly exceeded the cutoff of .05, the other goodness-of-fit indices indicated the 18 semantic differential items fit very well within the structure of these indices (χ^2= 37833, CFI = .97, NFI = .97, GFI = .95, AGFI = .93; Bentler, 1990; Chin & Todd, 1995; Gefen et al., 2000). Furthermore, the statistics of goodness-of-fit implied two subfactors of the eeriness index were robust enough to represent their own theoretical concepts (r = .44). In the practical work, two subfactors of the eeriness index could measure independently.

Table 8. Structural coefficients for the semantic items

	Humannes	Eeriness		Attractivenes
		Eerie	Spine-	
Inanimate–Living	.81	-	-	-
Synthetic–Real	.80	-	-	-
Mechanical Movement–Biological	.77	-	-	-
Human-made–Humanlike	.76	-	-	-
Without Definite Lifespan–Mortal	.67	-	-	-
Dull–Freaky[b]	-	.76	-	-
Predictable–Eerie[b]	-	.75	-	-
Plain–Weird[b]	-	.75	-	-
Ordinary–Supernatural	-	.66	-	-
Boring–Shocking	-	-	.77	-
Uninspiring–Spine-tingling	-	-	.72	-
Predictable–Thrilling	-	-	.65	-
Bland–Uncanny	-	-	.65	-
Unemotional–Hair-raising	-	-	.64	-
Ugly–Beautiful	-	-	-	.79
Repulsive–Agreeable	-	-	-	.78
Crude–Stylish	-	-	-	.77
Messy–Sleek	-	-	-	.69
Cronbach's α	.87	.82	.81	.85
Model	χ^2	df	GFI	AGFI
	3783	129	.95	.93
	NFI	CFI	RMR	RMSEA
	.97	.97	.15	0.061

[a] items sorted by the factor loading of each index

[b] new item candidate

The correlation analysis indicates the indices retained their construct validity (Table 9). In the final version, the attractiveness index had no significant correlation with eeriness (r=−.06, p=.069). The correlation of the attractiveness and eeriness indices with positive (vs. negative) affect was effectively eliminated. In addition, the eeriness index had no significant correlation with the humanness index (r=.04, p=.285).

Table 9. Correlation between attractiveness, eeriness, and humanness indices in the final version

	Attractiveness	Eeriness	Humanness
Attractiveness	-		
Eeriness	−.06 (p=.069)	-	
Humanness	.36 (p<.001)	.04 (p=.285)	-

Multidimensional scaling (MDS) was performed on the 18 semantic differential items. Figure 23 shows that the semantic differential items belonging to the humanness, eeriness, and attractiveness indices form three distinct, nonoverlapping subfactors. The four items belonging to the eerie subfactor and the five items belonging to the spine-tingling subfactor of the eeriness index were also separated (Table 8). These MDS results indicate the humanness, eeriness, and attractiveness indices could measure distinctly their corresponding concepts.

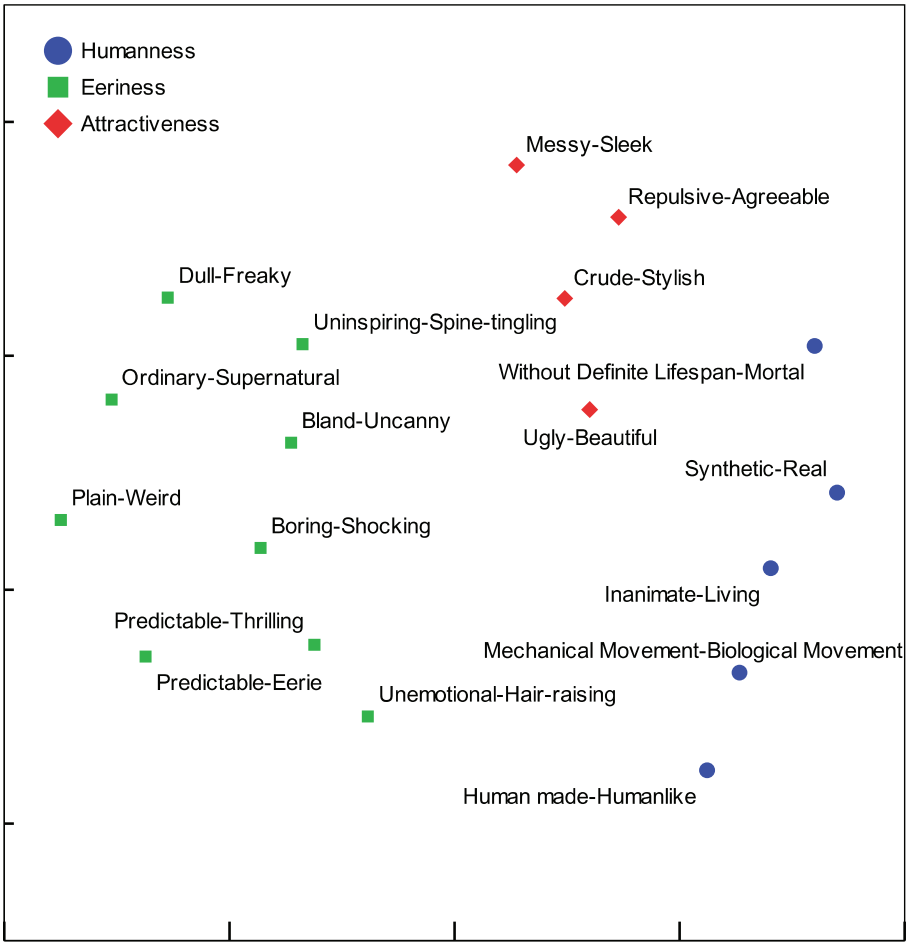

Figure 23. Multidimensional scaling of the 18 semantic differential items was performed based on participant ratings of the figures in the 12 video clips. Items from the humanness, eeriness, and attractiveness indices are widely separated.

The scatter plot showed that humanness and eeriness were decorrelated among various anthropomorphic characters (Figure 24). The insignificant correlation of the eeriness and humanness indices revealed that the final version of these indices had good discriminant validity and high reliability. The eeriness index also had an insignificant correlation with the humanness index (r=.04, p=.285). The attractiveness index yielded a high correlation with the humanness index (r=.36, p<.001), the data points vertically

aligned into three main groups: animation, robot, and human (Figure 25). Specifically, the results showed that the improved attractiveness and humanness indices were less affected by positive (vs. negative) affect than previously developed indices (Ho & MacDorman, 2010).

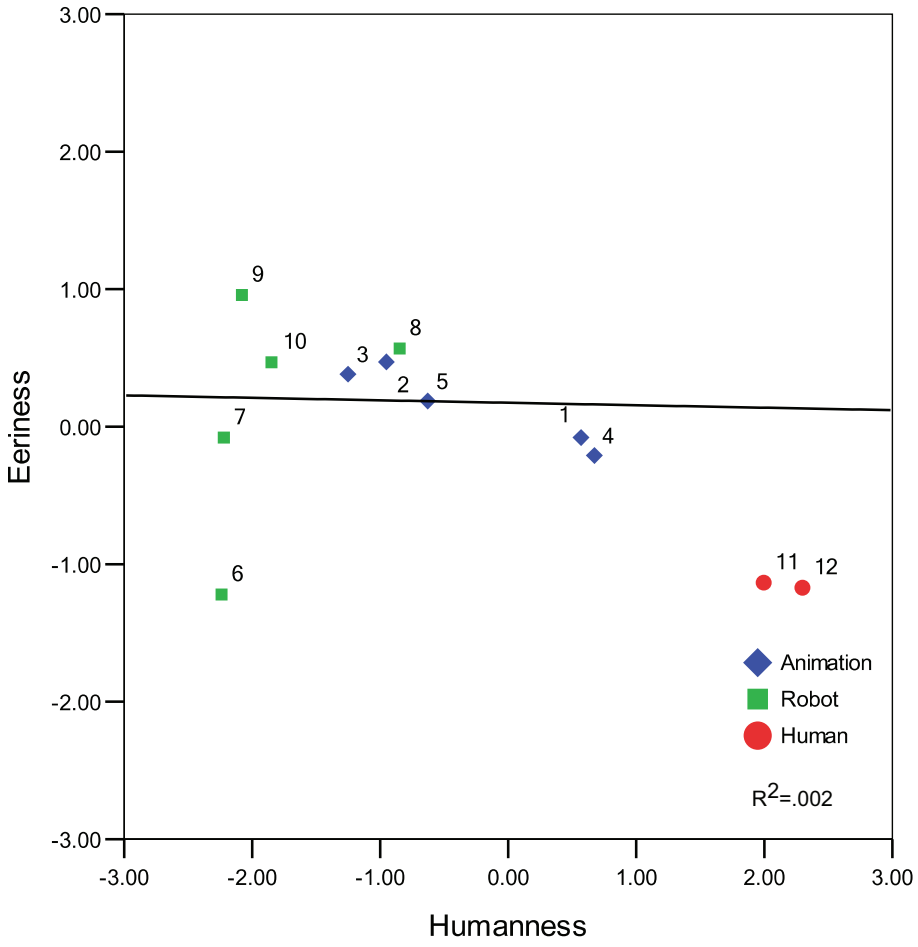

Figure 24. The final humanness and eeriness indices were not significantly correlated (r=.04, p=.285).

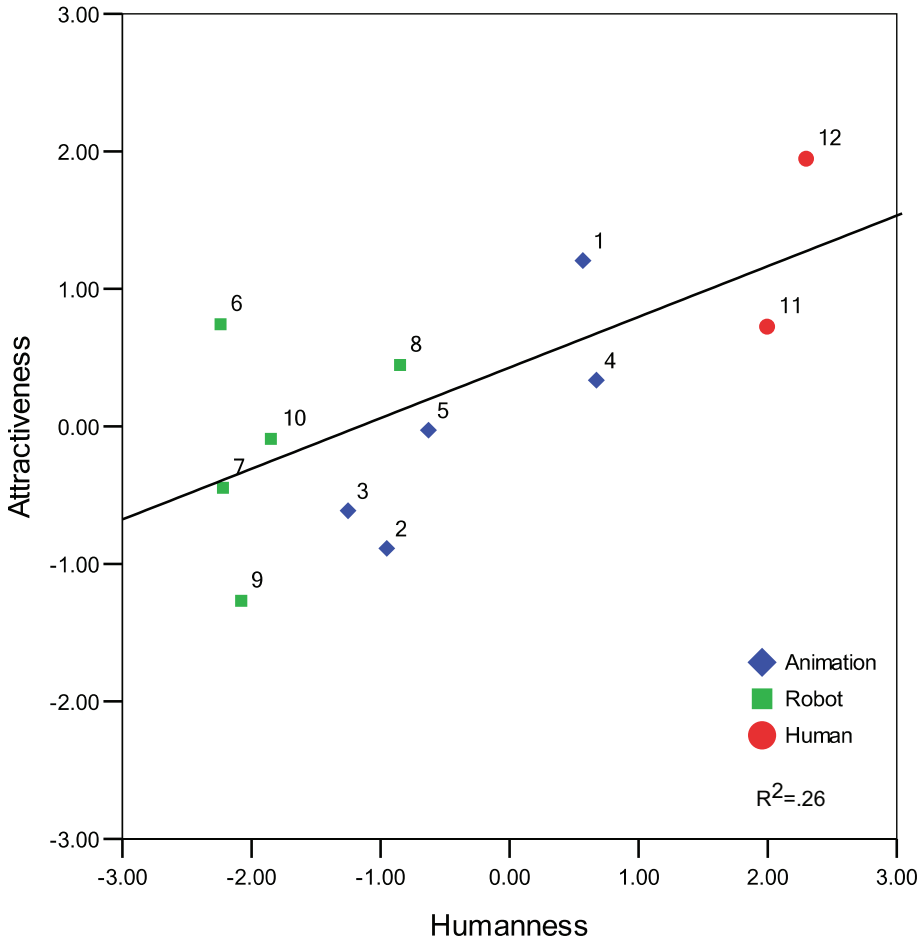

Figure 25. The humanness and attractiveness indices were significantly correlated but categorized into animation, robot, and human groups (r=.36, p<.001).

5. DISCUSSION

In the categorization exercise, the participants applied their schematic knowledge on the human categorization (Macrae & Bodenhausen, 2000). Yamauchi (2005) suggests that induction is carried out not just by matching similarity but also by abstract reasoning processes elicited by category information. In the laddering interview, many exclusively human characteristics were mentioned repeatedly. In addition, the automatic category activation is triggered when the robot category is identified; the robot category primes "machine," "simple work," and "human creation." The category activation comes through the heightened accessibility of material following the presentation of a priming stimulus (Devine, 1989). The facial expression is strong evidence of categorical boundaries based on the symmetrical associations of "Unconvincing Facial Expression" and "Mismatched Facial Expression" between the original defined animation and robot categories as well as "Convincing Facial Expression" between the self-identified android and human categories (cf. Looser & Wheatley, 2010). The participants considered the robots incapable of demonstrating proper facial expressions, whereas they believed that the animation characters performed appropriate facial expressions but only mismatch with the timing or related actions. In addition, participants firmly believed that only the humans could have genuine facial expressions. In addition, the eyes were the essential clue to determine whether the character looks human (Looser & Wheatley, 2010). The participants particularly care about eye movement. The participants' judgment about whether the character was convincing or unconvincing relied on the eyes.

Based on Ramey's assumption (2005), humans have difficulty categorizing androids or humanoids into such categories as "animate" or "inanimate," because they lie

at the boundary between these categories. Humans would repeatedly make the comparisons to solve the dilemma of cognitive dissonance to settle the uncertainty in concepts. Ramey thus considers the uncanny valley to be caused by stimuli at category boundaries, rather than a unique phenomenon related to anthropomorphic entities. However, the association results in this study do not support the assumption of Ramey. The participants still can give the android category with many monosemic items (e.g., contingency, mutual sense) rather than items with the related, multiple meanings (e.g., emotions, past personal experience). In addition, the android category found in this study may be close to the assumption of the third ontological category (Kahn et al., 2011, 2012; Kahn, Gary, & Shen, 2013). However, the reasons of the android category given by the adult participants are different from those given by the children. The differences might occur to the specific stage of psychological development. This issue remains for future work to clarify.

Considering the effect of the category, participants used less humanness, less eerie, and more attractive items to evaluate anthropomorphic entities. However, the self-evaluation of the category becomes the tool that can detect underestimating or overestimating biases during the category identification (Fox & Clemen, 2005). When the participants judge the anthropomorphic entities in terms of a continuous spectrum of human likeness, it is harder for them to determine how to partition human likeness. The participants underestimate merely humanlike robots with the fewer humanity traits. In addition, the participants would be influenced by their domain knowledge about anthropomorphism to tend to anchor their ignorance prior. The ordinary participants overestimate the automatic robots, which capable of finishing the simple task. Therefore,

these cognitive biases may have failed to reach conscious reflection (Arkes, 1991; Dunning et al., 2003; Kruger & Dunning, 1999; Pronin, 2007).

In the original association pattern technique (APT), the evaluated items must come from the pilot laddering interview (Gutman, 1982; Hofstede et al., 1998). However, it is limited in the small-scale studies because the evaluated items need to be tested. Applying the developed indices as the extraneous evaluation in the laddering interview will yield high content validity. Furthermore, the association models based on the indices' terms can be used to the convergent validity of laddering interview with respect to the content and structure of the means-end chains network that they reveal. Although the laddering interviews and evaluated terms have different data formats, the results of model testing indicate that both contain the same concepts. The terms used in the indices could serve as a snapshot of the relevant attributes, consequences, and values toward the anthropomorphic entities. In addition, the participant's prior categorization would facilitate the validity of the laddering interview.

Some new items of the revised indices came from the participants' own responses, such as dull, predictable, and weird. They might be more appropriate in modern English usage and provide better content validity than terms like "numbing" and "reassuring." The revised indices for anthropomorphic characters' attractiveness, eeriness, and humanness are shown to have high internal reliability. With respect to computer-animated human characters and robots, these indices demonstrate the bipolarity of the semantic space for assessing emotional responses and judgments of personality traits (Bentler, 1969; Gärling, 1976; Lorr & Wunderlich, 1988; Rosenberg et al., 1968; Van Schuur & Kiers, 1994). Confirmatory factor analysis was used to verify the theoretical

structure of these indices. Exploratory factor analysis demonstrates a comprehensive strategy for item selection prior to validation by confirmatory factor analysis (Gerbing & Hamilton, 1996). These indices appear to be valid for measuring their putative concepts. Compared with the original indices (Ho & MacDorman, 2010), the revised indices eliminate the categorical biases to measure independently. The two subscales of the eeriness index can serve as standalone measures to illustrate the perceived eeriness of the anthropomorphic characters. Relative to the animated characters, the robot entities had higher ratings in the eerie subscale but lower ratings in the spine-tingling subscale.

5.1. Limitations and Future Work

Considering laddering interviews, one of the limitations of APT is the oversimplified representation of the means-end chain network that considering the association linkages between concepts (i.e., the AC- and CV-linkages). Adding extra AA-, CC-, and VV-matrices that containing the same concepts in both rows and columns could lead to a the means-end chain network with a more comprehensive structure (Hofstede et al., 1998) In this study, the linkages between the attributes, consequences, and values are ignoring whether ladders are elicited from the same or different categories. Using nonlinear generalized canonical analysis (NGCA, Valette-Florence, 1998), kernel isometric mapping (ISOMAP), or other nonlinear dimensionality reduction techniques may not only help the researcher to identify the segments of the user's thought with specific means-end orientation, but also have the probability of the associations between the main means-end chain and any prespecified criterion, such as the participant-identified categories. In addition, the participant's emotional responses were kept in the

laddering transcription. The laddering interview involves a rationalization process; the reasons based on emotional responses became more relevant. For the future work, the emotional responses could be analyzed.

From the perspective of index development, there is considerable individual variation in emotional responses to humanoid robots and animated human characters. For example, although some participants were disturbed by the digital resurrection of the businessman Orville Redenbacher, other participants accepted the character as the real person. It is important to explore further the merely humanlike appearance that may influence the intensity of emotional responses. In addition, although the indices of Ho and MacDorman (2010) had high internal reliability and eliminated correlation with positive and negative affect, they might still be influenced by the effect of the category. When the user evaluated the interaction with the robot, the categorization process is activated simultaneously or even in advance. In other words, the predetermined category might be dominant. The participants might overestimate or underestimate the new items by the specific categories. The improved indices may need confirmation from a categorization task.

Although this study did not find age and gender to be significant factors in our population of undergraduates, these variables may be significant in a more heterogeneous sample that includes a broader range of ages. Past research has indicated that differences of culture and levels of exposure to robots can have a significant influence on attitudes (MacDorman, Vasudevan, & Ho, 2009). It is important to test the indices with different cultural populations.

It is also important to apply external criteria to assess the validity of the developed indices. For example, the microdynamics of interaction between an embodied agent and a human being can indicate the extent to which the human being is responding to the agent as if it were human (Cassell & Tartaro, 2007). The same information can also indicate an aversive response when the interaction breaks down. Nonverbal behavior, such as gaze frequency and duration, have been used to determine preference between still and computer-animated monkeys in experiments on the uncanny valley that used macaque monkeys as subjects (Steckenfinger & Ghazanfar, 2009), and similar methods have also been applied to human infants and adults in the study of attractiveness. Micro expressions, which convey emotional state, can be measured by optical motion tracking or electromyography. These kinds of behavioral metrics can be used to test the predictive validity of the developed indices, as can physiological variables, such as heart rate, respiration, and galvanic skin response, which can increase in response to fear, an emotion associated with uncanny stimuli (Ho et al., 2008). Functional magnetic resonance imaging (fMRI) can be used to correlate response strength on the indices with brain areas that have been identified with emotions associated with the uncanny valley (e.g., fear and anxiety in the central and lateral amygdale and medial hypothalamus, Panksepp, 2006; disgust in the anterior insular cortex and frontal operculum, Jabbi, Bastiaansen, & Keysers, 2008).

6. CONCLUSION

Although laddering interviews can uncover the underlying reasons for people's behaviors, only in combination with a categorization task will they reveal the "bias blind spots" (Pronin, 2002) that the participants assess overestimated and underestimated claims, which have been identified in the two comparisons: human versus robot, and robot versus animation. People tend to rely on introspective evidence despite the bias occurring nonconsciously. For instance, the automatic robot needs its software and hardware to work together flawlessly, but people devalue its performance. In other cases, people tend to convince themselves that their perceptions reflect reality though reality is less undesirable. For example, people still criticize the animation character, asserting that it cannot perform proper facial expressions because it is not a real human being, even through the character uses advanced motion capture to duplicate real human facial expressions. In this study, assessing the uncanny anthropomorphic characters not only elicits the eerie feeling but also produces the cognitive biases of the uncanny valley (e.g., facial expressions). It gives the researchers insight into human judgment (MacDorman & Ishiguro, 2006).

The improved set of uncanny valley indices confirms the measures for human perceptions of anthropomorphic characters that reliably assess relatively independent individual attitudes (Ho & MacDorman, 2010). Bartneck and colleagues (2009) note that developing indices for robots can benefit robot developers. However, the improved indices can also benefit animators. Comparing different characters and feature settings by means of the same index will help developers in making design decisions. The indices revised in this study have four advantages. First, they have excellent psychometric

properties. The theoretical structure keeps constant for both male and female participants and the large scale testing. Second, the internal reliability of the three indices are high. Third, the eeriness index, which could serve as the *y*-axis in Mori's graph, not only measures its named concept well but also is decorrelated from the *x*-axis, humanness as well as other contenders for the *y*-axis, the attractiveness and warmth indices.

The apparent independence of the humanness and eeriness indices enables anthropomorphic characters to be plotted along nearly orthogonal axes, as implied by Mori's (1970) original graph of the uncanny valley. Confirmatory factor analysis was used to verify the theoretical structure of the indices. The results indicate the development of robust instruments for the measurement of attractiveness, eeriness, and humanness. Fourth, the stimuli presented in this study were not limited to humanlike robots; they included computer-generated human characters. This study widens the range of stimuli to which the indices may be applied.

7. APPENDICES

7.1 IRB Statement

INDIANA UNIVERSITY

OFFICE OF RESEARCH ADMINISTRATION

Date: March 14, 2009

To: Dr. Karl MacDorman
Informatics
IT 487

From: Regina Wininger
Research Compliance Administration, IUPUI
UN 618

Subject: IUPUI/Clarian Institutional Review Committee - Exempt Review of Human Study

Study Number: EX0903-35B

Study Title: "Development of Attractiveness, Eeriness, and Human Indices"

Your application for approval of the study named above has been accepted as meeting the criteria of exempt research as described by Federal Regulations [45 CFR 46.101(b), paragraph 2]. A copy of the acceptance is enclosed for your file.

Although a continuing review is not required for an exempt study, prior approval must be obtained before change(s) to the originally approved study can be initiated. When you have completed your study, please inform our office in writing.

If the research is conducted at or funded by the VA, research may not be initiated until approval is received from the VA Research and Development Committee.

Please contact the Office of Health Care Billing and HIPAA Programs at 317-278-4891 for information regarding a Data Use Agreement, if applicable.

Enclosures: ☒ Copy of acceptance

Phone: 317-274-8289 • Fax: 317-274-5932 • Email: resrisk@iupui.edu • Website: http://research.iupui.edu

7.2 **Questionnaires**

Importance of items (3-point scale: Not important/Moderately important/Very important)

1. Perceived Humanness
    1. Artificial–Natural
    2. Synthetic–Real
    3. Inanimate–Living
    4. Human-made–Humanlike
    5. Mechanical Movement–Biological Movement
    6. Without Definite Lifespan–Mortal
2. Eeriness
    1. Reassuring–Eerie
    2. Numbing–Freaky
    3. Ordinary–Supernatural
    4. Uninspiring–Spine-tingling
    5. Boring–Shocking
    6. Predictable–Thrilling
    7. Bland–Uncanny
    8. Unemotional–Hair-raising
3. Attractiveness
    1. Unattractive–Attractive
    2. Ugly–Beautiful
    3. Repulsive–Agreeable

4. Crude–Stylish

5. Messy–Sleek

7.3 Published Article

Computers in Human Behavior 26 (2010) 1508–1518

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices

Chin-Chang Ho, Karl F. MacDorman *

Indiana University School of Informatics, 535 West Michigan Street, Indianapolis, IN 46202, USA

ARTICLE INFO

Article history:
Available online 8 June 2010

Keywords:
Affective appraisal
Embodied agents
Human–robot interaction
Psychometric scales
Social perception

ABSTRACT

Mori (1970) proposed a hypothetical graph describing a nonlinear relation between a character's degree of human likeness and the emotional response of the human perceiver. However, the index construction of these variables could result in their strong correlation, thus preventing rated characters from being plotted accurately. Phase 1 of this study tested the indices of the Godspeed questionnaire as measures of humanlike characters. The results indicate significant and strong correlations among the relevant indices (Bartneck, Kulić, Croft, & Zoghbi, 2009). Phase 2 of this study developed alternative indices with nonsignificant correlations ($p > .05$) between the proposed *y*-axis *eeriness* and *x*-axis *perceived humanness* ($r = .02$). The new *humanness* and *eeriness* indices facilitate plotting relations among rated characters of varying human likeness.

1. Plotting emotional responses to humanlike characters

Mori (1970) proposed a hypothetical graph describing a nonlinear relation between a character's degree of human likeness and the emotional response of the human perceiver (Fig. 1). The graph predicts that more human-looking characters will be perceived as more agreeable up to a point at which they become so human people find their nonhuman imperfections unsettling (MacDorman, Green, Ho, & Koch, 2009; MacDorman & Ishiguro, 2006; Mori, 1970). This dip in appraisal marks the start of the uncanny valley (*bukimi no tani* in Japanese). As characters near complete human likeness, they rise out of the valley, and people once again feel at ease with them. In essence, a character's imperfections expose a mismatch between the human qualities that are expected and the nonhuman qualities that instead follow, or vice versa. As an example of things that lie in the uncanny valley, Mori (1970) cites corpses, zombies, mannequins coming to life, and lifelike prosthetic hands.

Assuming the uncanny valley exists, what dependent variable is appropriate to represent Mori's graph? Mori referred to the *y*-axis as *shinwakan*, a neologism even in Japanese, which has been variously translated as familiarity, rapport, and comfort level. Bartneck, Kanda, Ishiguro, and Hagita (2009) have proposed using *likeability* to represent *shinwakan*, and they applied a *likeability* index to the evaluation of interactions with Ishiguro's android double, the Geminoid HI-1. Likeability is virtually synonymous with interpersonal warmth (Asch, 1946; Fiske, Cuddy, & Glick, 2007; Rosenberg, Nelson, & Vivekananthan, 1968), which is also strongly correlated with other important measures, such as comfortability, communality, sociability, and positive (vs. negative) affect (Abele & Wojciszke, 2007; MacDorman, Ough, & Ho, 2007; Mehrabian & Russell, 1974; Sproull, Subramani, Kiesler, Walker, & Waters, 1996; Wojciszke, Abele, & Baryla, 2009). Warmth is the primary dimension of human social perception, accounting for 53% of the variance in perceptions of everyday social behaviors (Fiske, Cuddy, Glick, & Xu, 2002; Fiske et al., 2007; Wojciszke, Bazinska, & Jaworski, 1998).

Despite the importance of warmth, this concept misses the essence of the uncanny valley. Mori (1970) refers to negative *shinwakan* as *bukimi*, which translates as eeriness. However, eeriness is not the negative anchor of warmth. A person can be cold and disagreeable without being eerie—at least not eerie in the way that an artificial human being is eerie. In addition, the set of negative emotions that predict eeriness (e.g., fear, anxiety, and disgust) are more specific than coldness (Ho, MacDorman, & Pramono, 2008). Thus, *shinwakan* and *bukimi* appear to constitute distinct dimensions.

Although much has been written on potential benchmarks for anthropomorphic robots (for reviews see Kahn et al., 2007; MacDorman & Cowley, 2006; MacDorman & Kahn, 2007), no indices have been developed and empirically validated for measuring *shinwakan* or related concepts across a range of humanlike stimuli, such as computer-animated human characters and humanoid robots. The Godspeed questionnaire, compiled by Bartneck, Kulić, Croft, and Zoghbi (2009), includes at least two concepts, *anthropomorphism* and *likeability*, that could potentially serve as the *x*- and *y*-axes of Mori's graph (Bartneck, Kanda, et al., 2009). Although the

* Corresponding author. Tel.: +1 317 215 7040.
E-mail address: kmacdorm@indiana.edu (K.F. MacDorman).
URL: http://www.macdorman.com (K.F. MacDorman).

0747-5632/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chb.2010.05.015

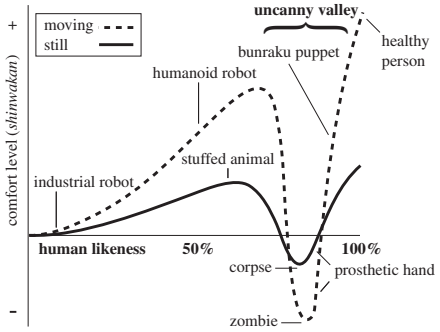

Fig. 1. Mori (1970) proposed a nonlinear relation, which is intensified by movement, between a character's degree of human likeness and the human perceiver's emotional response. The dip in emotional response just before total human likeness is referred to as the uncanny valley.

Godspeed questionnaire lists semantic differential items for each concept, the indices corresponding to these concepts have not been empirically tested as a group for overall reliability and validity. In addition, there is no index corresponding specifically to *eeriness*, a dimension that is arguably distinct from *likeability* but nevertheless important in determining whether a human-looking character has fallen into the uncanny valley.

Phase 1 of the current study evaluates the Godspeed indices based on participant ratings of computer-animated human characters and humanoid robots presented in video clips. The performance of the Godspeed indices in Phase 1 is used in Phase 2 to benchmark progress toward developing a new set of uncanny valley indices. The new set includes *eeriness* as a possible dimension for the *y*-axis in Mori's graph and decorrelates *eeriness* from *humanness* and *warmth*. Indices for *humanness*, *eeriness*, *warmth*, and *attractiveness* were developed in two rounds of testing using five methods of analysis: (1) adjectives that could serve as potential anchors for semantic differential items were selected for each index and rated on their positive (vs. negative) affect, and inversely correlated adjectives that had similar affective ratings were paired in semantic differential items; (2) reliability analysis was used to remove less reliable items from each index; (3) exploratory factor analysis was used to determine the geometric solution of the indices by oblique rotation; (4) correlation analysis was used to decorrelate the indices from interpersonal warmth; and (5) confirmatory factor analysis was used to test their theoretical structure.

2. An empirical analysis of the Godspeed indices

Bartneck, Kulić, et al. (2009) assembled five indices composed of semantic differential items in the Godspeed questionnaire to assist developers in creating embodied social agents. The indices are *anthropomorphism* (Powers & Kiesler, 2006), *animacy* (converted from Likert scales; Lee, Park, & Song, 2005), *likeability* (Monahan, 1998), *perceived intelligence* (Warner & Sugarman, 1996), and *perceived safety* (Kulić & Croft, 2007). The purpose of Phase 1 of this study is twofold: to test for the first time the validity, reliability, and theoretical structure of these indices *as a set* for a range of robots and computer-animated human characters and, specifically, to determine whether *anthropomorphism* and *likeability* are sufficiently decorrelated to serve as *x*- and *y*-axes in plotting people's emotional response to characters that vary in their degree of perceived human likeness. It should be noted that in the past development of these indices, no attempt had been made to decorrelate them from positive (vs. negative) affect or from each other. As an example of this, *anthropomorphism* and *animacy* have a semantic differential item in common, *artificial–lifelike*.

Several of the indices, including *anthropomorphism*, would appear to be correlated with positive (vs. negative) affect, interpersonal warmth, and *likeability*, based on the face validity of the opposing anchors used for their semantic differential items. For example, *fake*, *moving rigidly*, and other anchors used to indicate low anthropomorphism have a negative nuance compared to *natural*, *moving elegantly*, and other anchors used to indicate high anthropomorphism. This trend continues for *animacy* with low animacy anchors like *dead*, *stagnant*, and *apathetic* and high animacy anchors like *alive*, *lively*, and *responsive*; for *perceived intelligence* with low intelligence anchors like *ignorant*, *foolish*, and *irresponsible* and high intelligence anchors like *knowledgeable*, *sensible*, and *responsible;* and for *perceived safety* with low safety anchors like *agitated* and *anxious* and high safety anchors like *calm* and *relaxed*.

Given that interpersonal warmth is the dominant dimension of human social perception and the apparent alignment of the anchors with positive and negative affect, a general concern is that each of the Godspeed indices may not measure the concept after which it was named but instead measures some convolution of that concept and interpersonal warmth. A more specific concern for our study is that, if *anthropomorphism* and *likeability* are strongly correlated, a scatter plot of characters rated along these axes will be highly skewed (Fig. 2). The plot will not accurately depict the characters' scores on the convoluted variable, and topological relations will be distorted.

2.1. Research methods

2.1.1. Participants

Participants were recruited from a list of randomly selected undergraduate students and recent graduates of a nine-campus Midwestern university. Among the 384 participants, 161 (41.9%) were male and 223 (58.1%) were female, 187 (48.7%) were under 20 years old, 162 (42.2%) were 21 to 25 years old, and 35 (9.1%) were over 26 years old. The participants reflected the demographics of the university's undergraduate population (80.1% non-Hispanic white, 6.9% African-American, 3.4% Asian, 3.0% Hispanic, and 6.6% foreign or unclassified). With respect to the sample's representativeness of the undergraduate population as a whole, the measurement error range was ±5.0% at a 95% confidence level.

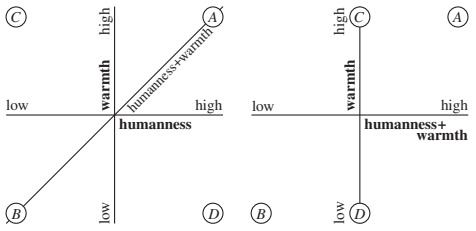

Fig. 2. Plotting an index that is a composite of two or more dimensions on a single axis distorts topological relations among observations. To illustrate this, four characters, labeled A, B, C, and D, are plotted against the *humanness* and *warmth* axes for the graph on the left and the *humanness + warmth* and *warmth* axes for the graph on the right. For the graph on the right, the degree of *humanness* of the low humanness character C and the high humanness character D cannot be distinguished. In addition, C is closer to A than B, and D is closer to B than A, although the distances should be equal.

There were no significant differences among the studies reported in this paper by gender or age.

2.1.2. Materials and procedures

Each participant viewed 10 video clips presented one at a time in random order (see Fig. 3). There were five video clips of three-dimensional computer-animated characters and five of robots. The video clips were displayed using a width of 480 pixels and a height of 360 pixels, which is a 4:3 aspect ratio. The clips were 15–30 s in length. Clips were played in a continuous loop while participants answered a survey on the figure featured in each video clip.

The survey consisted of the Godspeed questionnaire, which is composed of five indices and 24 semantic differential items. The *anthropomorphism* index has five items, the *animacy* index has six items, the *likeability* index has five items, the *perceived intelligence* index has five items, and the *perceived safety* index has three items (Table 1).

2.1.3. Statistical analysis

Cronbach's α was used to measure the reliability of each index. Confirmatory factor analysis was used to verify whether the 24 semantic differential items divide into five factors corresponding to the five Godspeed indices. If the results of confirmatory factor analysis were inconsistent with the construct dimensions, the items could not represent the concepts of the indices. In addition, correlation analysis was used to evaluate the relation among the indices and to test their discriminant validity. Multidimensional scaling (MDS) was used to create a (Euclidean) distance matrix for all pairs of the 24 semantic differential items to approximate their distance from each other in a space that has been reduced

Fig. 3. The five video clips on the top row contain computer-animated human characters from the films (1) *Final Fantasy: The Spirits Within*, (2) *The Incredibles*, and (3) *The Polar Express*, (4) an Orville Redenbacher popcorn advertisement, and (5) a technology demonstration of the *Heavy Rain* video game. The remaining five video clips contain (6) iRobot's Roomba 570, (7) JSK Laboratory's Kotaro, (8) Hanson Robotics's Elvis and (9) Eva, and (10) Le Trung's Aiko.

Table 1
Structural coefficients for the Godspeed indices.

Items[a]	Anthropomorphism	Animacy	Likeability	Perceived intelligence	Perceived safety
Machinelike–Humanlike	.89	–	–	–	–
Artificial–Lifelike	.87	–	–	–	–
Fake–Natural	.85	–	–	–	–
Unconscious–Conscious	.76	–	–	–	–
Moving rigidly–Moving elegantly	.76	–	–	–	–
Mechanical–Organic	–	.88	–	–	–
Artificial–Lifelike	–	.87	–	–	–
Dead–Alive	–	.79	–	–	–
Stagnant–Lively	–	.64	–	–	–
Apathetic–Responsive	–	.59	–	–	–
Inert–Interactive	–	.57	–	–	–
Awful–Nice	–	–	.86	–	–
Unpleasant–Pleasant	–	–	.85	–	–
Dislike–Like	–	–	.83	–	–
Unfriendly–Friendly	–	–	.81	–	–
Unkind–Kind	–	–	.81	–	–
Ignorant–Knowledgeable	–	–	–	.81	–
Unintelligent–Intelligent	–	–	–	.79	–
Incompetent–Competent	–	–	–	.78	–
Foolish–Sensible	–	–	–	.74	–
Irresponsible–Responsible	–	–	–	.70	–
Agitated–Calm	–	–	–	–	.84
Anxious–Relaxed	–	–	–	–	.70
Surprised–Quiescent	–	–	–	–	.19
Cronbach's α	.91	.88	.92	.87	.60
Model	χ^2	df	GFI	AGFI	
	3927.25	242	.86	.82	
	NFI	CFI	RMR	RMSEA	
	.98	.98	.086	.088	

[a] Items are sorted by the factor loading of each index.

from 24 to 2 dimensions. The distance matrix was used to visualize similarities and dissimilarities among the items. Internal reliability and correlation analysis were performed using SPSS, confirmatory factor analysis was performed using LISREL, and multidimensional scaling was performed using MATLAB.

2.2. Results

To confirm the reliability and the validity of the Godspeed indices, an internal reliability test was conducted. The results showed that the *likeability* and *anthropomorphism* indices had the highest reliability with a Cronbach's α of .92 and .91, respectively. The Cronbach's α of *animacy* and *perceived intelligence* was .88 and .87, respectively. However, *perceived safety* had low reliability with a Cronbach's α of .60, which is below the standard .70 cutoff (Nunnally, 1978).

Confirmatory factor analysis was used to test the theoretical structure of the Godspeed indices. Table 1 shows the factor loadings of the 24 semantic differential items. In the model, two goodness-of-fit indices (RMR = .086; RMSEA = .088) exceeded the standard .05 cutoff, indicating that the 24 semantic differential items did not fit well in the structure of these five indices (χ^2 = 3927.25, CFI = .98, NFI = .98, GFI = .86, AGFI = 0.82; Bentler, 1990; Chin & Todd, 1995; Gefen, Straub, & Boudreau, 2000). A serious problem was that several factor loadings could not reach a high level, such as *stagnant–lively*, *inert–interactive*, and *apathetic–responsive* for *animacy* and *surprised–quiescent* for *perceived safety*. The result is that the latent constructs could not capture more than half their variances.

Another serious problem was the significant and extremely high correlation between *anthropomorphism*, *likeability*, *animacy*, and *perceived intelligence* (Table 2). The correlations ranged from .67 for *anthropomorphism* and *perceived intelligence* to .89 for *anthropomorphism* and *animacy*. This suggests that those concepts had no discriminant validity. In other words, they were all measuring the same concept instead of measuring distinct concepts.

Multidimensional scaling was performed on the 24 semantic differential items. Fig. 4 shows that semantic differential items belonging to the *anthropomorphism* and *animacy* indices are distributed across a large overlapping region. Although the *likeability* items are packed closely together, they are wholly contained within the region circumscribed by the *anthropomorphism* and *animacy* items. The MDS results indicate that the *anthropomorphism*, *animacy*, and *likeability* indices are unable to measure distinctly their corresponding concepts.

The conclusion that the Godspeed indices lack discriminant validity is further supported by the fact that the spread of data points in a scatter plot followed a diagonal line of humanness: all the robots were located in the lower-left area, and the computer-animated human characters were located in the upper-right area (Figs. 5–7). *Likeability* was significantly (p = .000) and highly correlated with *anthropomorphism* (r = .73), *animacy* (r = .74), and *perceived intelligence* (r = .71). These findings indicate that the Godspeed indices could not measure the intended concepts independently of positive (vs. negative) affect. In addition, the *anthropomorphism* index could not separate the robots by their degree of humanness despite a nonanthropomorphic robot, Roomba 570, being included in the group.

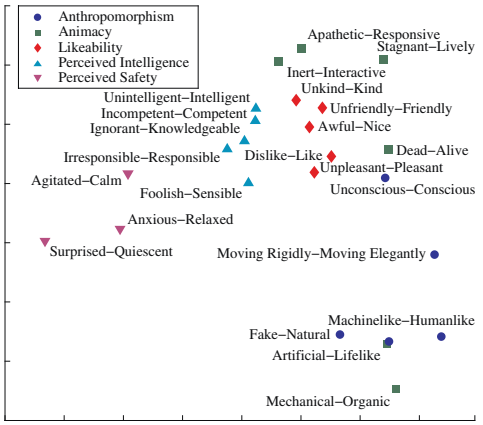

Fig. 4. Multidimensional scaling of the 24 semantic differential items was performed based on participant ratings of the figures in the 10 video clips. Items from the *anthropomorphism* and *animacy* indices are spread out across a large overlapping region, which includes the *likeability* items.

3. The development of humanness, warmth, eeriness, and attractiveness indices

The results of Phase 1 of this study found that the Godspeed indices did not represent their concepts independently of positive (vs. negative) affect. Hence, in Phase 2 an alternative set of indices is developed to measure participants' attitudes toward anthropomorphic characters: *perceived humanness*, *warmth*, *eeriness*, and *attractiveness*.

The first three indices are motivated by the original graph of the uncanny valley proposed by Mori (1970). Studies on the uncanny valley typically manipulate as an independent variable a character's "objective" humanness—the human photorealism of the character's morphology, skin texture, motion quality, or other formal property (MacDorman, Coram, Ho, & Patel, 2010; MacDorman et al., 2009; Seyama & Nagayama, 2007). However, it is also useful to have a corresponding measure of its subjective or perceived humanness to check whether the objective manipulation is having the intended effect. Interpersonal warmth is useful to include, because it is the dominant dimension of human social perception and strongly correlated with concepts identified with *shinwakan*, the *y*-axis of Mori's graph, such as comfort level, likeability, and rapport.

Table 2
Correlation between anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety.

	Anthropomorphism	Animacy	Likeability	Perceived intelligence	Perceived safety
Anthropomorphism	–				
Animacy	.89***	–			
Likeability	.73***	.74***	–		
Perceived Intelligence	.67***	.72***	.71***	–	
Perceived Safety	.06**	–.01	.20***	.17***	–

** p < .01 (2-tailed).
*** p < .001 (2-tailed).

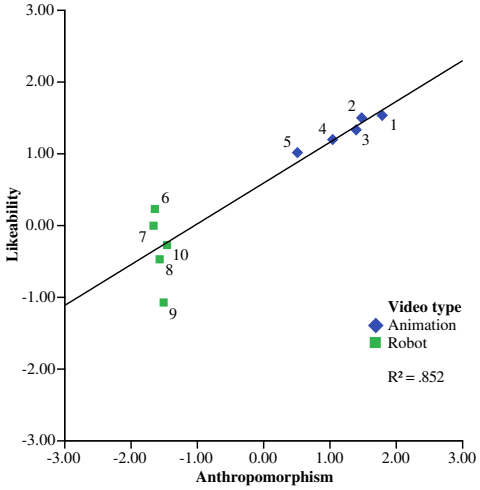

Fig. 5. The *anthropomorphism* and *likeability* indices of the Godspeed questionnaire are significantly and strongly correlated ($p = .000$, $r = .73$). The ratings of the computer-animated human characters are nearly collinear, as are the ratings of the robots. The *anthropomorphism* index is unable to discriminate the robots by their degree of humanness. The humanoid robot, Kotaro, was rated as having slightly lower *anthropomorphism* than the nonanthropomorphic robot, Roomba 570.

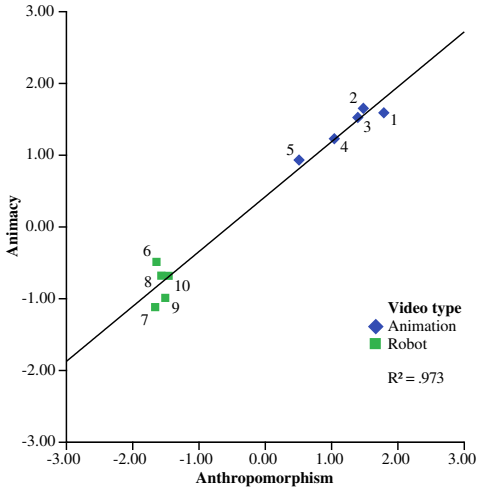

Fig. 6. The *anthropomorphism* and *animacy* indices of the Godspeed questionnaire are significantly and strongly correlated ($p = .000$, $r = .89$). This indicates they may be measuring the same concept. The ratings of the computer-animated human characters are nearly collinear, and there is little separation among the ratings of the robots.

Eeriness, which is conceptually distinct from negative warmth (i.e., interpersonal coldness), would need to be included in any set of indices on the uncanny valley, as it corresponds to the phenomenon to be explained.

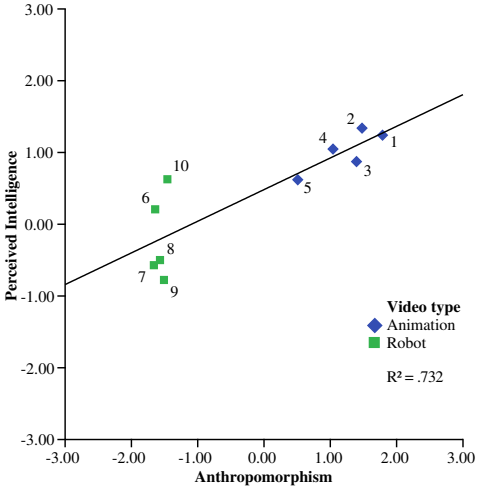

Fig. 7. The *anthropomorphism* and *perceived intelligence* indices of the Godspeed questionnaire are significantly and strongly correlated ($p = .000$, $r = .67$).

An *attractiveness* index is included, because physical attractiveness is an important dimension in explanations of the uncanny valley based on evolved perceptual and cognitive mechanisms for mate selection and pathogen avoidance (MacDorman & Ishiguro, 2006; MacDorman et al., 2009). Bilateral symmetry, clear skin, certain proportions of the face and body, and other observable markers of attractiveness are correlated with reproductive fitness as measured by a range of physiological variables, including sperm count, strength of female orgasm, hormonal and immune system levels, and the ability to conceive (Jasienska, Ziomkiewicz, Ellison, Lipson, & Thune, 2004; Jones, Little, & Perrett, 2004; Manning, Scutt, & Lewis-Jones, 1998; Thornhill & Gangestad, 1993; Thornhill, Gangestad, & Comer, 1995). There is an extensive literature exploring the evolutionary and cultural basis for perceptions of attractiveness and their pervasive impact on human behavior (Cunningham, Roberts, Barbee, Druen, & Wu, 1995; Jones, 1995; Langlois et al., 1987; Langlois et al., 2000). Attractiveness is known to influence many kinds of decisions, even without principled reasons, including decisions of moral consequence (Cunningham, 1986). Therefore, it is important to control for the effects of attractiveness in studies on the uncanny valley.

3.1. Research goal

The goal of Phase 2 of this study is to develop valid and reliable indices for *perceived humanness*, *warmth*, *eeriness*, and *attractiveness* based on corresponding semantic differential items, such that *perceived humanness* and *eeriness* are not significantly correlated with each other or with *warmth* or *attractiveness*. The naïve development of *perceived humanness* and *eeriness* indices could confound these dimensions with interpersonal warmth. If *eeriness*, for example, were strongly correlated with interpersonal warmth, wicked but artfully rendered villains might be rated eerier than amiable but uncanny-looking heroes (e.g., the queen in Walt Disney's 1937 hand-animated film *Snow White* versus the conductor in Robert Zemeckis's 2004 computer-animated film *The Polar Express*). Such an index would not be able to detect characters that

had fallen into the uncanny valley as described by Mori (1970). In this study, decorrelation between indices was achieved for *eeriness* but only partly achieved for *perceived humanness*.

Semantic differential items were used in Phase 2, because they can reduce acquiescence bias (i.e., the tendency of participants to agree with statements) without lowering psychometric quality (Friborg, Martinussen, & Rosenvinge, 2006; Lorr & Wunderlich, 1988). To decorrelate the *humanness*, *eeriness*, and *attractiveness* indices from interpersonal warmth, the opponent adjective pairs of their semantic differential items went through a process of selection to find adjectives that have about the same level of positive (vs. negative) affect. These adjectives are paired in semantic differential scales so the indices that accumulate their values are not correlated with positive (vs. negative) affect. In addition, this study attempts to adhere to the following guidelines in constructing *humanness*, *eeriness*, and *attractiveness* indices: (1) the opponent adjective pairs should be moderately or strongly inversely correlated; (2) items corresponding to a single, unidimensional concept should load on the same factor when applying exploratory factor analysis as a heuristic tool for index development (Comrey, 1978); (3) the positive and negative anchors of eeriness and humanness adjective pairs should be nearly uncorrelated with the *warmth* or *pleasure* indices, and the attractiveness item pairs should have at most a medium correlation; (4) there should be at least three semantic differential scales per index to enable the estimation of reliability; and (5) the reliability of the indices should be acceptable (Cronbach's $\alpha \geqslant .70$).

3.2. Methods

3.2.1. Participants

In the initial round of testing, there were 19 participants, 13 (68.4%) male and 6 (31.6%) female, of whom 7 (36.8%) were 21–25 years old, 4 (21.1%) were 26–30, 5 (26.3%) were 31–35, and 3 (15.8%) were over 36. Most participants were human–computer interaction (HCI) graduate students, young professionals, and HCI-related professionals.

In the second round of testing, participants were recruited from a random selection of undergraduate students and recent graduates of a nine-campus Midwestern university. Among the 253 participants, 112 (44.3%) were male and 141 (55.7%) were female, 216 (85.4%) were under 25 years old, 20 (7.9%) were 26–30, and 17 (6.7%) were over 31. The participants reflected the demographics of the university's undergraduate population. The measurement error range was ±6.16% at a 95% confidence level.

3.2.2. Materials and procedures

The video clips and method of presentation were the same as in the previous study. Each participant viewed 10 video clips presented one at a time in random order (see Fig. 3). There were five video clips of three-dimensional computer-animated characters and five of robots. The video clips were displayed using a width of 480 pixels and a height of 360 pixels, which is a 4:3 aspect ratio. Most clips were 15–30 s in length. Clips were played in a continuous loop while participants answered a survey on the figure featured in each video clip. The initial round of the survey consisted of 22 semantic differential items: seven from the *perceived humanness* index, eight from the *eeriness* index, and seven from the *attractiveness* index. The second round of the survey consisted of 29 semantic differential items: 10 from the *humanness* index, 8 from the *eeriness* index, and 11 from the *attractiveness* index.

3.2.3. Statistical analysis

Internal reliability was used to measure how reliable items were for their indices in each round of testing. Exploratory factor analysis, which applied the principal components analysis method and the Promax rotation, was used to verify that the semantic differential items loaded on factors corresponding to their named concepts. In addition, *artificial–natural* in the *humanness* index, *reassuring–eerie* in the *eeriness* index, and *unattractive–attractive* in the *attractiveness* index were chosen as "sanity check" items to verify the correctness of indices. A sanity check item has high face validity but does not necessarily meet the other criteria for an item, such as being correlated with interpersonal warmth. If the results of factor analysis varied from the sanity check's dimension and showed low factor loadings, new items should be developed and added to the index in the next round. Correlation analysis showed the relation between indices and verified the discriminant validity of indices during testing. Confirmatory factor analysis was used to verify the theoretical structure of the new set of uncanny valley indices. Finally, multidimensional scaling was used to visualize similarities and dissimilarities among the semantic differential items by reducing the dimensionality of the space from 19 to 2 dimensions. Internal reliability, exploratory factor analysis, and correlation analysis were performed using SPSS, confirmatory factor analysis was performed using LISREL, and multidimensional scaling was performed using MATLAB.

3.3. Results

3.3.1. Humanness index

A pool of seven items was initially selected for the *humanness* index (see Table 3). *Artificial–natural* was the sanity check for the *humanness* index. The overall internal reliability of the initial test was relatively high (Cronbach $\alpha = .85$). The initial exploratory factor analysis with no iterations showed all items loaded on a single factor that explained 57.33% of the variance. The reliability was improved by removing *genderless–male or female*, *uncommunicative–bigmouthed*, and *automatic–deliberate*.

These items were replaced with *inanimate–living, mechanical movement–biological movement*, and *synthetic–real* in the second round of testing. The internal reliability in the second round of testing remained the same. As with the initial round of testing, exploratory factor analysis extracted (with no iterations) one major factor that explained 60.79% of the variance. However, the newly added items contributed higher factor loadings than those of *genderless–male or female*, *uncommunicative–bigmouthed*, and *automatic–deliberate*.

In the final version of the index, *artificial–natural*, *human-made–humanlike*, *without definite lifespan–mortal*, *inanimate–living*,

Table 3
Reliability and factor loadings of the humanness index.

Items[a]	Round 1	Round 2	Final
Artificial–Natural[b]	.83	.87	.90
Human-made–Humanlike	.82	.85	.88
Innocent of Morals–Aware of Right and Wrong[d]	.82	.77	–
Without Definite Lifespan–Mortal	.81	.84	.85
Genderless–Male or Female[d]	.71	.63	–
Uncommunicative–Bigmouthed[d]	.66	.62	–
Automatic–Deliberate[d]	.62	.52	–
Inanimate–Living[c]	–	.86	.88
Mechanical Movement–Biological Movement[c]	–	.86	.86
Synthetic–Real[c]	–	.86	.90
Total variance explained	57.33%	60.79%	68.96%
Cronbach's α	.85	.85	.92

[a] Items are sorted by the factor loading of the initial round of testing.
[b] The sanity check.
[c] Items added in the second round of testing.
[d] Items excluded from the final version.

mechanical movement–biological movement, and *synthetic–real* were the measurement items. Therefore, the final version of the humanness index would retain six items. Its internal reliability was high (Cronbach's $\alpha = .92$), and it explained 68.96% of the variance.

3.3.2. Eeriness index

A pool of eight items was initially selected for the *eeriness* index (see Table 4). *Reassuring–eerie* was the sanity check for the *eeriness* index. The overall internal reliability in the initial round of testing was .80. The initial exploratory factor analysis with three iterations showed that two major factors were extracted. *Reassuring–eerie*, *numbering–freaky*, *bland–uncanny*, and *ordinary–supernatural* loaded on the first factor, which explained 43.42% of the variance. The internal reliability of the first factor was .76. *Unemotional–hair-raising*, *uninspiring–spine-tingling*, *boring–shocking*, and *predicable–thrilling* loaded on the second factor, which explained 19.80% of the variance. The internal reliability of the second factor was .79.

Because the initial results met the reliability criterion, the second round of testing was followed by exploratory factor analysis to check whether the items represented the eeriness index appropriately. Although the internal reliability of the second round of data was .74, the exploratory factor analysis result with three iterations was similar to the initial testing. *Unemotional–hair-raising*, *uninspiring–spine-tingling*, *boring–shocking*, *predicable–thrilling*, and *bland–uncanny* loaded on the first dimension, which explained 38.40% of the variance. *Reassuring–eerie*, *numbering–freaky*, and *ordinary–supernatural* loaded on the second dimension, which explained 22.93% of the variance.

Because the two dimensions explained sufficient variance and were both relevant to the concept of eeriness, all items in the *eeriness* index were retained in the final version. For follow-up confirmatory factor analysis, the factor corresponding to the *reassuring–eerie*, *numbering–freaky*, and *ordinary–supernatural* items was referred to as *eerie*, and its internal reliability was .71; the factor corresponding to the *unemotional–hair-raising*, *uninspiring–spine-tingling*, *boring–shocking*, *predicable–thrilling*, and *bland–uncanny* items was referred to as *spine-tingling*, and its internal reliability was .81. Therefore, the final version of the *eeriness* index would retain eight items that explained 62.04% of the variance and held an overall internal reliability of .74.

Table 4
Reliability and factor loadings of the eeriness index.

Items[a]	Round 1		Round 2		Final	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Reassuring–Eerie[b]	.91	−.34	−.22	.87	−.22	.87
Numbing–Freaky	.80	.06	.05	.82	.05	.82
Ordinary–Supernatural	.68	.13	.20	.67	.20	.67
Bland–Uncanny	.68	.16	.70	.09	.70	.09
Unemotional–Hair-raising	−.14	.85	.75	−.23	.75	−.23
Uninspiring–Spine-tingling	.05	.82	.78	.08	.78	.08
Predictable–Thrilling	−.08	.75	.76	−.09	.76	−.09
Boring–Shocking	.32	.66	.77	.17	.77	.17
Total variance explained	43.42%	19.80%	38.40%	22.93%	38.40%	22.93%
Cronbach's α	.76	.79	.81	.71	.81	.71
Overall Cronbach's α	.80		.74		.74	

[a] Items are sorted by the factor loading of the initial round of testing.
[b] The sanity check.

3.3.3. Attractiveness index

A pool of seven items was initially selected for the *attractiveness* index (see Table 5). Opponent adjectives that were rated as having similar levels of positive (vs. negative) affect were paired in semantic differential items. *Unattractive–attractive* was the sanity check for the *attractiveness* index. The initial internal reliability was .78. The initial exploratory factor analysis with three iterations extracted two major factors. *Unpretentious–alluring*, *prim–eye-catching*, *modest–sensual*, *unadorned–showy*, and *plain-featured–racy* loaded on the first factor, which explained 44.09% of the variance. Only *homely–slick* was grouped with *unattractive–attractive* in the second factor, which explained 14.83% of the variance.

The initial result's first factor did not contain *unattractive–attractive* and thus did not appear to be measuring attractiveness. Therefore, four items were added in the second round of testing: *ugly–beautiful*, *repulsive–agreeable*, *crude–stylish*, and *messy–sleek*. The internal reliability of the data in the second round of testing was .84. Although exploratory factor analysis extracted two factors in three iterations, the four newly added items loaded on the same factor as *unattractive–attractive*, and this factor explained 39.75% of the variance. The cronbach's α of these five items was .90. The final version of the *attractiveness* index would retain these five items, which explained 70.93% of the variance. Although these items had high reliability and face validity, the opponent adjectives did not have the same level of positive (vs. negative) affect. Thus, the items would be unlikely to meet the goal of decorrelating *attractiveness* from *warmth*.

3.3.4. Pleasure and warmth indices

Sad–happy, *bad–good*, *terrible–wonderful*, and *annoyed–pleased* comprised the *pleasure* index in the initial round of testing. The internal reliability of the *pleasure* index was acceptable (Cronbach's $\alpha = .79$). The *pleasure* index was used to assess the correlations among indices. If the *attractiveness*, *humanness*, and *eeriness* indices correlated highly with the *pleasure* index, it means that the positive (vs. negative) affect in these indices might dilute their discriminant validity. *Cold-hearted–warm-hearted*, *hostile–friendly*, *spiteful–well-intentioned*, *ill-tempered–good-natured*, and *grumpy–cheerful* comprised the *warmth* index in the second round of testing. The internal reliability of the *warmth* index was high (Cronbach's

Table 5
Reliability and factor loadings of the attractiveness index.

Items[a]	Round 1		Round 2		Final
	Factor 1	Factor 2	Factor 1	Factor 2	
Unpretentious–Alluring[d]	.75	.07	.22	.57	–
Modest–Sensual[d]	.75	.02	−.10	.70	–
Plain-featured–Racy[d]	.74	−.05	−.09	.77	–
Unadorned–Showy[d]	.73	−.05	−.03	.71	–
Prim–Eye-catching[d]	.73	−.01	.07	.62	–
Homely–Slick[d]	−.15	.92	.35	.26	–
Unattractive–Attractive[b]	.21	.69	.84	.05	.87
Repulsive–Agreeable[c]	–	–	.88	−.18	.82
Ugly–Beautiful[c]	–	–	.86	.04	.88
Messy–Sleek[c]	–	–	.81	−.04	.79
Crude–Stylish[c]	–	–	.80	.06	.82
Total variance explained	44.09%	14.83%	39.75%	16.32%	70.93%
Cronbach's α	.79	.49	.87	.72	.90
Overall Cronbach's α	.78		.84		.90

[a] Items are sorted by the factor loading of the initial round of testing.
[b] The sanity check.
[c] Items added in the second round of testing.
[d] Items excluded from the final version.

α = .88). Like the *pleasure* index, the *warmth* index in the second round of testing was designed to assess its correlation with other indices. If any index showed a high correlation with the *warmth* index, its items should be modified to eliminate this correlation.

3.3.5. Validation of the final version of the indices

Based on two rounds of testing, five items were constructed for the final version of the *attractiveness* index, eight items were constructed for the *eeriness* index, and six items were constructed for the *humanness* index (Tables 3–5). Confirmatory factor analysis was used to test the theoretical structure of the final set. Table 6 shows the factor loadings for the 19 semantic differential items of the final set. Although one goodness-of-fit index (RMSEA = .075) slightly exceeded the cutoff of .05, the other goodness-of-fit indices indicated that the 19 semantic differential items fit moderately well within the structure of these indices (χ^2 = 1229.29, CFI = .97, NFI = .97, GFI = .91, AGFI = 0.88; Bentler, 1990; Chin & Todd, 1995; Gefen et al., 2000).

The correlation analysis indicated that the indices retained their construct validity (Table 7). In the final version, the *attractiveness* index had no significant correlation with *eeriness* (r = −.03, p = .316). The correlation of the *attractiveness* and *eeriness* indices with positive (vs. negative) affect was effectively eliminated. In addition, the *eeriness* index had no significant correlation with the *humanness* index (r = .02, p = .514).

Multidimensional scaling was performed on the 19 semantic differential items. Fig. 8 shows that semantic differential items belonging to the *humanness*, *eeriness*, and *attractiveness* indices are in three distinct, nonoverlapping regions. The three items belonging to the *eerie* subfactor and the five items belonging to the *spine-tingling* subfactor of the *eeriness* index (listed in Table 6) are also widely separated. These MDS results indicate that the *perceived humanness*, *eeriness*, and *attractiveness* indices can measure distinctly their corresponding concepts.

Table 6
Structural coefficients for the semantic differential items.

Items[a]	Perceived Humanness	Eeriness		Attractiveness
		Eerie	Spine-tingling	
Artificial–Natural	.89	–	–	–
Synthetic–Real	.87	–	–	–
Inanimate–Living	.86	–	–	–
Human-made–Humanlike	.84	–	–	–
Mechanical Movement–Biological Movement	.83	–	–	–
Without Definite Lifespan–Mortal	.80	–	–	–
Reassuring–Eerie	–	.79	–	–
Numbing–Freaky	–	.69	–	–
Ordinary–Supernatural	–	.55	–	–
Uninspiring–Spine-tingling	–	–	.75	–
Boring–Shocking	–	–	.75	–
Predictable–Thrilling	–	–	.66	–
Bland–Uncanny	–	–	.63	–
Unemotional–Hair-raising	–	–	.63	–
Unattractive–Attractive	–	–	–	.87
Ugly–Beautiful	–	–	–	.87
Repulsive–Agreeable	–	–	–	.78
Crude–Stylish	–	–	–	.75
Messy–Sleek	–	–	–	.69
Cronbach's α	.92	.71	.81	.90
Model	χ^2	df	GFI	AGFI
	1229.29	146	.91	.88
	NFI	CFI	RMR	RMSEA
	.97	.97	.23	.075

[a] Items sorted by the factor loading of each index.

Table 7
Correlation between the attractiveness, eeriness, humanness, and warmth indices in the final version.

	Attractiveness	Eeriness	Humanness	Warmth
Attractiveness	–			
Eeriness	−.03	–		
Humanness	.61***	.02	–	
Warmth	.62***	−.05	.66***	–

*** p < .001 (2-tailed).

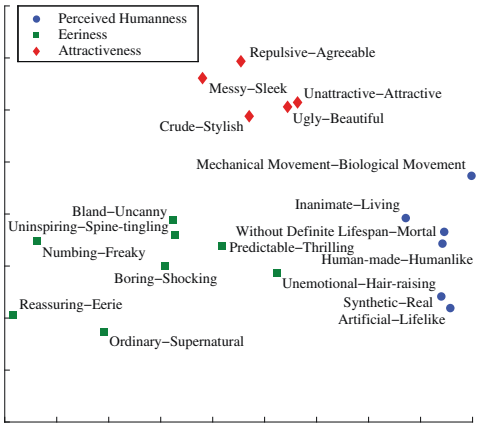

Fig. 8. Multidimensional scaling of the 19 semantic differential items was performed based on participant ratings of the figures in the 10 video clips. Items from the *perceived humanness*, *eeriness*, and *attractiveness* indices are widely separated.

The scatter plot shows that *humanness* and *eeriness* were decorrelated (Fig. 9), and *warmth* and *eeriness* were also decorrelated (Fig. 10). The data points did not follow a diagonal line as they had in the Godspeed indices. The insignificant correlation of the *eeriness* and *humanness* indices revealed that the final version of these indices could have good discriminant validity and high reliability. The *eeriness* index also had an insignificant correlation with the *warmth* index (r = −.05, p = .083). Although the attractiveness index yielded a high correlation with the *humanness* index (r = .61, p = .000), the data points vertically aligned into two main groups. Specifically this analysis showed that the *attractiveness* and *humanness* indices were somewhat less affected by positive (vs. negative) affect than *anthropomorphism* in the Godspeed indices. Although the *humanness* index was not correlated with the *eeriness* index after two rounds of testing, the *humanness* index maintained a high correlation with the *warmth* index (r = .66, p = .000). This analysis indicated that the notion of *warmth* might strongly overlap with the concept of *humanness* in practical circumstances. It is difficult to obtain discriminant validity; however, this may be improved in future studies.

4. Discussion

In Phase 1 of this study, the results of the validity analysis identified several problems with the Godspeed indices. The reliability

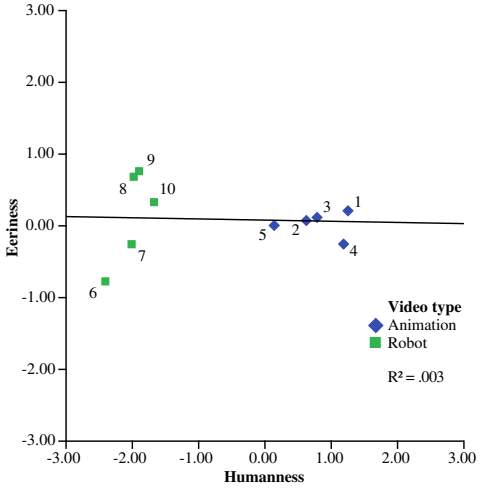

Fig. 9. The developed *humanness* and *eeriness* indices are not significantly correlated ($p = .514$, $r = .02$).

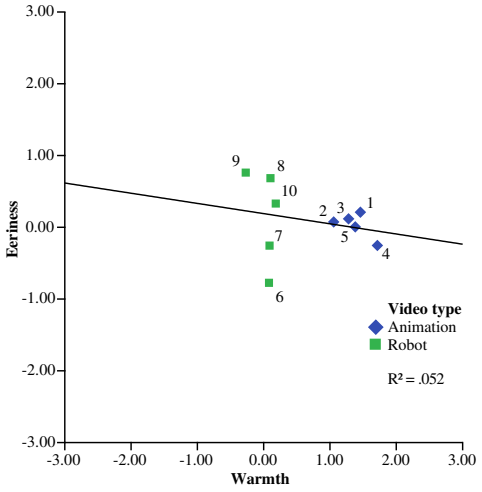

Fig. 10. The developed *warmth* and *eeriness* indices are not significantly correlated ($p = .083$, $r = -.05$).

of *perceived safety* was below the standard .70 cutoff. Confirmatory factor analysis also found inconsistencies in these indices and indicated that several items should be removed. However, the most serious problem was that *anthropomorphism*, *animacy*, *likeability*, and *perceived intelligence* were highly correlated with each other. This correlation indicates that they may be measuring the same concept, not separate concepts. These findings indicate the Godspeed indices are not appropriate as distinct concepts for evaluating anthropomorphic agents.

Therefore, Phase 2 included a new set of uncanny valley indices. After two rounds of testing, the developed indices for anthropomorphic characters' *attractiveness*, *eeriness*, and *humanness* were shown to have high internal reliability. With respect to computer-animated human characters and robots, these indices demonstrate the bipolarity of the semantic space for assessing people's emotional responses and judgments of personality traits (Bentler, 1969; Gärling, 1976; Lorr & Wunderlich, 1988; Rosenberg et al., 1968; Van Schuur & Kiers, 1994). Exploratory factor analysis was used to determine which items were retained for each index, and confirmatory factor analysis was used to verify the theoretical structure of the indices. Exploratory factor analysis demonstrated a comprehensive strategy for model selection prior to the validation by confirmatory factor analysis (Gerbing & Hamilton, 1996). In general, these indices appear to be valid for measuring their putative concepts.

4.1. Limitations and future work

The new indices were developed and validated with a particular set of stimuli, but it is important to retest them with other sets of stimuli. A limitation of the current set is that there were more nonhuman characteristics in the humanoid robots than in the animated human characters. To increase the variation within each group, less polished animations should be included, such as those rendered by video game software engines, and more polished human-looking robots should also be included, such as the Geminoid F developed by Hiroshi Ishiguro's laboratory at Osaka University and Kokoro Co. Ltd.

There is also considerable individual variation in emotional responses to humanoid robots and animated human characters. For example, although some participants were disturbed by the digital resurrection of the businessman Orville Redenbacher, other participants accepted the character as the real person. It is important to explore demographic factors that may influence the intensity of emotional responses. Although our study did not find age and gender to be significant factors in our population of undergraduates, these participant variables may be significant in a more heterogeneous sample that includes a broader range of ages. Past research has indicated that differences of culture and levels of exposure to robots can have a significant influence on attitudes (MacDorman, Vasudevan, & Ho, 2009). It is important to test the indices with different populations.

It is also important to apply external criteria to assess the validity of the developed indices. For example, the microdynamics of interaction between an embodied agent and a human being can indicate the extent to which the human being is responding to the agent as if it were human (Cassell & Tartaro, 2007). The same information can also indicate an aversive response when the interaction breaks down. Nonverbal behavior, such as gaze frequency and duration, have been used to determine preference between still and computer-animated monkeys in experiments on the uncanny valley that used macaque monkeys as subjects (Steckenfinger & Ghazanfar, 2009), and similar methods have also been applied to human infants and adults in the study of attractiveness. Facial expressions, which convey emotional state, can be measured by optical motion tracking or electromyography. These kinds of behavioral metrics can be used to test the predictive validity of the developed indices, as can physiological variables, such as heart rate, respiration, and galvanic skin response, which can increase in response to fear, an emotion associated with uncanny stimuli (Ho et al., 2008). Functional magnetic resonance imaging (fMRI) can be used to correlate response strength on the indices with brain areas that have been identified with emotions associated with the uncanny valley (e.g., fear and anxiety in the central and lateral amygdala

and medial hypothalamus, Panksepp, 2006; disgust in the anterior insular cortex and frontal operculum; Jabbi, Bastiaansen, & Keysers, 2008).

5. Conclusion

The set of uncanny valley indices developed in the current study are new measures for human perceptions of anthropomorphic characters that reliably assess four relatively independent individual attitudes. Bartneck, Kulić, et al. (2009) note that developing indices for robots can benefit robot developers. Comparing different robots and robot settings by means of the same index will help developers in making design decisions. The indices developed in this study have four advantages. First, they have excellent psychometric properties. The factor structure remains constant for both male and female participants and across two rounds of testing. Second, the internal reliability of the four indices is high. Third, the *eeriness* index, which could serve as the *y*-axis in Mori's graph, not only measures its named concept well but also is decorrelated from the *humanness*, *warmth*, and *attractiveness* indices. The apparent independence of the *humanness* and *eeriness* indices enables anthropomorphic characters to be plotted along nearly orthogonal axes, as implied by Mori's (1970) original graph of the uncanny valley. Confirmatory factor analysis was used to verify the theoretical structure of the indices. The results indicate the development of robust instruments for the dimensions of *attractiveness*, *eeriness*, *humanness*, and *warmth*. Fourth, the stimuli presented in this study were not limited to humanlike robots; they included computer-generated human characters. This widens the range of stimuli to which the indices may be applied.

Acknowledgments

The authors would like to express their gratitude to Himalaya Patel, Wade Mitchell, and the anonymous reviewers for their thoughtful suggestions for improving this paper. The IUPUI/Clarian Research Compliance Administration has approved this study (EX0903-35B). This study was supported by an IUPUI Signature Center grant.

References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus other. *Journal of Personality and Social Psychology, 93*(5), 751–763.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*(3), 259–290.

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009). My robotic doppelganger: A critical look at the uncanny valley theory. In *Proceedings of the 18th IEEE international symposium on robot and human interactive communication* (pp. 269–276). Toyama, Japan.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81.

Bentler, P. M. (1969). Semantic space is (approximately) bipolar. *Journal of Psychology, 71*(1), 33–40.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 170*(2), 238–246.

Cassell, J., & Tartaro, A. (2007). Intersubjectivity in human–agent interaction. *Interaction Studies, 8*(3), 391–410.

Chin, W. W., & Todd, P. A. (1995). On the use, usefulness, and ease of use of structural equation modeling in MIS research: A note of caution. *MIS Quarterly, 19*(2), 237–246.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46*(4), 648–659.

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi- experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology, 50*(5), 925–935.

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). "Their ideas of beauty are on the whole the same as ours?" Consistency and variability in cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology, 68*(2), 261–279.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

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.

Gärling, T. (1976). A multidimensional scaling and semantic differential technique study of the perception of environmental settings. *Scandinavian Journal of Psychology, 17*(1), 323–332.

Gefen, D., Straub, D., & Boudreau, M. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems, 4*(7), 1–79.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*(1), 62–72.

Ho, C.-C., MacDorman, K., & Pramono, Z. A. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings. In *Proceedings of the third ACM/IEEE international conference on human–robot interaction* (pp. 169–176). March 11–14, Amsterdam, The Netherlands.

Jabbi, M., Bastiaansen, J., & Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS ONE, 3*(8), e2939.

Jasienska, G., Ziomkiewicz, A., Ellison, P., Lipson, S., & Thune, I. (2004). Large breasts and narrow waists indicate high reproductive potential in women. *Proceedings of the Royal Society of London: Biological Sciences, 271*(1545), 1213–1217.

Jones, D. (1995). Sexual selection, physical attractiveness, and facial neoteny: Cross-cultural evidence and implications. *Current Anthropology, 36*(5), 723–748.

Jones, B. C., Little, A. C., & Perrett, D. I. (2004). When facial attractiveness is only skin deep. *Perception, 33*(5), 569–576.

Kahn, P. H., Jr., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., et al. (2007). What is a human? Toward psychological benchmarks in the field of human–robot interaction. *Interaction Studies, 8*(3), 363–390.

Kulić, D., & Croft, E. (2007). Physiological and subjective responses to articulated robot motion. *Robotica, 25*, 13–27.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*(3), 390–423.

Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., & Jenkins, V. Y. (1987). Infant preferences for attractive faces: Rudiments of a stereotype. *Developmental Psychology, 23*(3), 363–369.

Lee, K. M., Park, N., & Song, H. (2005). Can a robot be perceived as a developing creature? *Human Communication Research, 31*(4), 538–563.

Lorr, M., & Wunderlich, R. A. (1988). A semantic differential mood scale. *Journal of Clinical Psychology, 44*(1), 33–36.

MacDorman, K. F. & Cowley, S. J. (2006). Long-term relationships as a benchmark for robot personhood. In *Proceedings of the 15th IEEE international symposium on robot and human interactive communication* (pp. 378–383). September 6–9, Hatfield, UK.

MacDorman, K. F., Coram, J. A., Ho, C.-C., & Patel, H. (2010). Gender differences in the impact of presentational factors in human character animation on decisions of ethical consequence. *Presence: Teleoperators and Virtual Environments, 19*(3).

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. (2009). Too real for comfort: Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695–710.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies, 7*(3), 297–337.

MacDorman, K. F., & Kahn, P. H. Jr., (2007). Introduction to the special issue on psychological benchmarks of human–robot interaction. *Interaction Studies, 8*(3), 359–362.

MacDorman, K. F., Ough, S., & Ho, C.-C. (2007). Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research, 36*(4), 283–301.

MacDorman, K. F., Vasudevan, S. K., & Ho, C.-C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society, 23*(4), 485–510.

Manning, J. T., Scutt, D., & Lewis-Jones, D. I. (1998). Developmental stability, ejaculate size and sperm quality in men. *Evolution and Human Behavior, 19*(5), 273–282.

Mehrabian, A., & Russell, J. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Monahan, J. L. (1998). I don't know you but I like you: The effects of nonconscious affect on person perception. *Human Communication Research, 24*, 480–500.

Mori, M. (1970). *Bukimi no tani* (the uncanny valley). *Energy, 7*(4), 33–35.

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

Panksepp, J. (2006). Emotional endophenotypes in evolutionary psychiatry. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 30*(5), 774–784.

Powers, A. & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proceedings of the first ACM SIGCHI/SIGART conference on human–robot interaction* (pp. 218–225). March 2–3, Salt Lake City, Utah, USA.

Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*(4), 283–294.

Seyama, J., & Nagayama, R. S. (2007). The uncanny valley: The effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments, 16*(4), 337–351.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human–Computer Interaction, 11*(2), 97–124.

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences, 106*(43), 18362–18366.

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. *Human Nature, 4*(3), 237–269.

Thornhill, R., Gangestad, S. W., & Comer, R. (1995). Human female orgasm and mate fluctuating asymmetry. *Animal Behaviour, 50*(6), 1601–1615.

Van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts and what model to use instead. *Applied Psychological Measurement, 18*(2), 97–110.

Warner, R. M., & Sugarman, D. B. (1996). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology, 50*, 792–799.

Wojciszke, B., Abele, A. E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology, 39*(6), 973–990.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1245–1257.

REFERENCES

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus other. *Journal of Personality and Social Psychology, 93*(5), 751–763.

Acton, G. S., & Revelle, W. (2002). Interpersonal personality measures show circumplex structure based on new psychometric criteria. *Journal of Personality Assessment, 97*(3), 446–471.

Altarriba, J., & Bauer, L. M. (2004). The distinctiveness of emotion concepts: A comparison between emotion, abstract, and concrete words. *American Journal of Psychology, 117*(3), 389–410.

Arkes, H. R. 1991. Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110*(3), 486–498.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*(3), 259–290.

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009). My Robotic Doppelganger- A Critical Look at the Uncanny Valley. *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 269–276), Sept. 27–Oct. 2, Toyama, Japan.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics, 1*(1)*,* 71–81.

Becker-Asano, C., & Ishiguro, H. (2011). Evaluating facial displays of emotion for the android robot Geminoid F. *Proceedings of IEEE SSCI Workshop on Affective Computational Intelligence* (pp. 22–29), April 11–15, Paris, France.

Becker-Asano, C., Ogawa, K., Nishio, S., & Ishiguro, H. (2010). Exploring the Uncanny Valley with Geminoid HI-1 in a real-world application. *Proceedings of IADIS International Conference Interfaces and Human Computer Interaction* (pp. 121–128), July 26–30, Freiburg, German.

Bentler, P. M. (1969). Semantic space is (approximately) bipolar. *Journal of Psychology, 71*(1), 33–40.

Bødker, S. (2006). When second wave HCI meets third wave challenges. *Proceedings of the 4th Nordic Conference on Human-computer Interaction* (NordiCHI '06) (pp. 1–8), New York, NY.

Bulter, M., & Joschko, L. (2007). Final Fantasy or the Incredibles: Ultra-realistic animation, aesthetic engagement and the uncanny valley. *Animation Studies, 3*, 55–63.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, *115*(3), 401–423.

Capra, M. G. (2005). Factor Analysis of Card Sort Data: An Alternative to Hierarchical Cluster Analysis. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, (pp. 691–95), Santa Monica, CA. Human Factors and Ergonomics Society.

Chin, W. W., & Todd, P. A. (1995). On the use, usefulness, and ease of use of structural equation modeling in MIS research: A note of caution. *MIS Quarterly, 19*(2), 237–246.

Claeys, C., Swinnen, P., & Vanden Abeele, P. (1995). Consumer's means-end chains for "think" and "feel" products. *International Journal of Research in Marketing, 12*(3), 193–208.

Comrey, A. L. (1978). Common methodological problem in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46*(4), 648–659.

Cortes, B. P., Demoulin, S., Rodriguez, R. T., Rodriguez, A. P., & Leyens, J. P. (2005). Infrahumanization or familiarity? Attribution of uniquely human emotions to the self, the ingroup, and the outgroup. *Personality and Social Psychology Bulletin, 31*(2), 243–253.

Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press. Cambridge, United Kingdom.

Curtis, V., Aunger, R., & Rabie, T. (2004). Evidence that disgust evolved to protect from risk of disease. *Proceedings of the Royal Society of London: Biological Sciences, 271*(Suppl. 4), S131-S133.

Dautenhahn K. (2007). Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 679–704.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18.

Dickstein, R., & Mills, V. (2000). Usability testing at the University of Arizona library: How to let the users in on the design. *Information Technology and Libraries,19*(3), 144–151.

DiSalvo, C., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). All robots are not created equal: The design and perception of humanoid robot heads. *Proceeding of Designing Interactive* Systems (pp. 321–326), London, United Kingdom.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87.

Evertt, M. G., & Borgatti, S. P. (1999). The centrality of groups and classes. *Journal of Mathematical Sociology, 23*(3), 181–201.

Evertt, M. G., & Borgatti, S. P. (2005). Extending centrality. In P. J. Carrington, J. Scott, & S. Wasserman (Eds), *Models and Methods in Social Network Analysis* (pp. 57–76). New York, NY: Cambridge University Press.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77—83.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Floridi, I., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Fong, T., Nourbakhsh,I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*, 143–166.

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science, 51*(9), 1417–1432.

Freedman, Y. (2012). Is it real… or is it motion capture? The battle to redefine animation in the age of digital performance. *The Velvet Light Trap, 69,* 38–49.

Freud, S. (1916-1917[1915]). Trauer und Melancholie, *Intern. Zschr. ärztl. Psychoanal*, *4*, 277–287; Mourning and melancholia. Standard Edition, 14, 243–258.

Freud, S. (1919/2003). The uncanny [das Unheimliche] (D. McLintock, Trans.). New York, NY: Penguin.

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.

Gärling, T. (1976). A multidimensional scaling and semantic differential technique study of the perception of environmental settings. *Scandinavian Journal of Psychology, 17*(1), 323–332.

Gefen, D., Straub, D., & Boudreau, M. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems 4* (7), 1–79.

Gerard, H. B., & Mathewson, G. C. (1966). The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology, 2*(3), 278–287.

Gerbing, D. W. & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*(1), 62–72.

Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, *1*(3), 203–275.

Gueguen, N., & De Gail, M. A. (2003). The effect of smiling on helping behavior: Smiling and good Samaritan behavior. *Communication Reports, 16*, 133–140.

Gutman, J. (1977). Uncovering the distinctions people make versus the use of multiattribute model: Do a number of little truths make wisdom? *Proceedings of the 23th Annual Conference of the Advertising Research Foundation* (pages 71–76). New York, NY.

Gutman, J. (1982). A means–end chain model based on consumer categorization processes. *Journal of Marketing, 46*(2), 60–72.

Gutman, J. (1991). Exploring the nature of linkages between consequences and values. *Journal of Business Research, 22*(2), 143–148.

Harnad, S. (1987). Introduction: Psychological and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 1–25). New York, NY: Cambridge University Press.

Harnad, S. (1987). Category induction and representation. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 535–565). New York, NY: Cambridge University Press.

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, *26*(6), 1508–1518.

Ho, C.-C., MacDorman, K. F., & Pramono, Z. A. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings. Proceedings of the Third ACM/IEEE International Conference on Human-Robot Interaction (pp. 169–176), March 11-14. Amsterdam, Netherlands.

Hofstede, F., Audenaert, A., Steenkamp, J-B E.M., & Wedel, M. (1998). An investigation into the association pattern techniques as a quantitative approach to measuring means-end chains. *International Journal of Research in Marketing, 15*(1), 37–50.

International Federation of Robotics (2013). *Considerable increase of medical robots and logistic systems.* Retrieved from http://www.ifr.org/news/ifr-press-release/considerable-increase-of-medical-robots-and-logistic-systems-552/

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, 21, 384–388.

Iwamura, Y., Shiomi, M., Kanda, T. Ishiguro, H., & Hagita, N. (2011). Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets? *Proceedings of the 6th ACM/IEEE International Conference on Human-robot Interaction* (HRI '11) (pp. 449–456), Lausanne, Switzerland.

Jans, G., & Calvi, L. (2006). Using laddering and association techniques to develop a user-friendly mobile (city) application. In R. Meersman, Z. Tari, P. Herrero et al. (Eds.), *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops* (pp. 1956–1965). Berlin, Germany: Springer.

Jentsch, E. (1906). Zur Psychologie des Unheimlichen (On the psychology of the uncanny), *Psychiatrisch-Neurologische Wochenschrift, 8*(22), 195–198.

Joule, R. V., & Azdia, T. (2003). Cognitive dissonance, double forced compliance, and commitment. *European Journal of Social Psychology*, *33*(4), 565–571.

Kahn, P. H., Gary, H. E., & Shen, S. (2013). Children's social relationships with current and near-future robots. *Child Development Perspectives*, *7*(1), 32–37.

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., & Shen, S. (2012). “Robovie, you'll have to go into the closet now”: Children’s social and moral relationships with a humanoid robot. *Developmental Psychology*, *48*(2), 303–314.

Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J. H., & Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (HRI '11) (pp. 159–160). Lausanne, Switzerland.

Kanda,T., Nishio, S., Ishiguro, H., & Hagita, N. (2009). Interactive humanoid robots and androids in children’s lives. *Children, Youth and Environments, 19*(1), 12–33.

Kikutani, M., Roberson, D., & Hanley, J. R. (2010). Categorical perception for unfamiliar faces: The effect of covert and overt face learning. *Psychological Science, 21*(6), 865–872.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Lane, R. D., Chua, P., & Dolan, R. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing pictures. *Neuropsychologia, 37*(9)*,* 989–997.

Larsen, R. J., & Diener, E. (1992). Problems and promises with the circumplex model of emotion. *Review of Personality and Social Psychology, 13,* 25–59.

Liotti, M., Mayberg, H. S., Brannan, S. K., McGinnis, S., Jerabek, P., & Fox, P. T. (2000). Differential limbic-cortical correlates of sadness and anxiety in healthy subjects: Implications for affective disorders. *Biological Psychiatry, 48*(1), 30–42.

Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science, 21*(12), 1854–1862.

Lorr, M. & Wunderlich, R. A. (1988). A semantic differential mood scale. *Journal of Clinical Psychology, 44*(1), 33–36.

Ludemann, P. & Nelson, C. A. (1988). The categorical representation of facial expressions by 7-month-old infants. *Developmental Psychology, 24*(4), 492–501.

MacDorman, K. F. & Cowley, S. J. (2006). Long-term relationships as a benchmark for robot personhood. *In Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 378–383). September 6-9, Hatfield, United Kingdom.

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. (2009). Too real for comfort: Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3)*,* 695–710.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. Interaction Studies, 7(3), 297–337.

MacDorman, K. F., & Kahn, P. H., Jr. (2007). Introduction to the special issue on psychological benchmarks of human-robot interaction. *Interaction Studies, 8*(3), 359–362.

MacDorman, K. F., Ough, S., & Ho, C.-C. (2007). Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. Journal of New Music Research, 36(4), 283–301.

Malhotra, Y., & Galletta, D. F. (1999). Extending the technology acceptance model to account for social influence: Theoretical bases and empirical validation. *Proceedings of the 32nd Hawaii International Conference on System Sciences* (HICSS 32, vol. 1), January 5–8, Hawaii, USA.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*(1), 93–120.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–768.

Mehrabian, A., & Russell, J. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Mitchell, W. J., Ho, C.-C., Patel, H., & MacDorman, K. F. (2011). Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior, 27*(1), 402–412.

Mitchell, W. J., Szerszen, Sr., K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception, 2*(1), 10–12.

Mori, M. (1970/2012). Bukimi no tani [the uncanny valley]. (K. F. MacDorman, Trans) *Energy*, *7*(4), 33–35.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological Review*, *98*(2), 164–181.

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2004). Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication* (pp. 35–40), September 20–22, Kurashiki, Okayama Japan.

Nomura, T., Shintani, T., Fujii, K., & Hokabe, K. (2007). Experimental investigation of relationships between anxiety, negative attitudes, and allowable distance of robots. *Proceedings of the Second IASTED International Conference on Human–Computer Interaction* (pp. 13–18), March 14–16, Chamonix, France.

Northoff, G., Richter, A., Gressner, M., Schlagenhauf, F., Stephan, K., Fell, J., Baumgart, F., Kaulisch, T., Kötter, R., Leschinger, A. Bargel, B., Witzel, T. Hinrichs, H., Bogerts, B., Scheich, H., & Heinze, H.-J. (2000). Functional dissociation between medial and lateral prefrontal cortical spatiotemporal activation in negative and positive emotions: A combined fMRI/MEG study. *Cerebral Cortex, 10*(1), 93–107.

O'Reilly, C., & Chatman, J. (1986). Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization on prosocial behavior. *Journal of Applied Psychology, 71*(3), 492–499.

Paradiso, S., Johnson, D. L., Andreasen, N. C., O'Leary, D. S., Watkins, G. L., Ponto, L. L. B., & Hichwa, R. D. (1999). Cerebral blood flow changes associated with attribution of emotional valence to pleasant, unpleasant, and neutral visual stimuli in a PET study of normal subjects. *American Journal of Psychiatry, 156*(10)*,* 1618–1629.

Philips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., Bullmore, E. T., Perrett, D. I., Rowland, D., Williams, S. C. R., Gray, J. A., & David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*(6650), 495–498.

Plantec, P. (2007). The digital eye: Crossing the great uncanny valley. (Dec. 19, 2007). Retrieved from http://www.awn.com/articles/production/crossing-great-uncanny-valley

Plantec, P. (2008). The digital eye: Image metrics attempts to leap the uncanny valley. (August 7, 2008). Retrieved from http://www.awn.com/articles/technology/digital-eye-image-metrics-attempts-leap-uncanny-valley

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 197–219). Hillsdale, NJ: Erlbaum.

Powers, A., Kiesler, S., & Goetz, J. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *Human-Computer Interaction Institute*, 105.

Prinz, J. J. (2004). Gut Reactions: A perceptual theory of emotion. New York, NY: Oxford University Press.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, *11*(1), 37–43.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381.

Provine, R. R. (2000). *Laughter: A scientific investigation.* New York, NY: Penguin .

Rafaeli, E., & Revelle, W. (2006). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion, 30*(1), 1–12.

Ramey, C. H. (2005). The uncanny valley of similarities concerning abortion, baldness, heaps of sand, and humanlike robots. *Proceedings of the Views of the Uncanny Valley Workshop, IEEE-RAS International Conference on Humanoid Robots* (pp. 8–13). December 5, Tsukuba, Japan.

Ramey, C. H. (2006). An inventory of reported characteristics for home computers, robots, and human beings: Applications for android science and the uncanny valley. *Proceedings of the ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science* (pp. 21–25). July 26, Vancouver, Canada.

Reynolds, T.J., & Gutman, J. (1988) Laddering theory, method, analysis, and interpretation. *Journal of Advertising Research, 28*(1), 11–31.

Rhodes, G. & Zebrowitz, L. A. (Eds.) (2001). Facial attractiveness: Evolutionary, cognitive, and social perspectives. Westport, CT: Ablex Publishing.

Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*(4)*,* 283–294.

Rugg, G., Eva, M., Mahmood, A., Rehman, N., Andrews, S., & Davies, S. (2002). Eliciting information about organizational culture via laddering. *Information Systems Journal, 12*(3), 215–229.

Rugg, G., & McGeorge, P. (1997). The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems, 12*(4), 80–93.

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology*, *37*(3), 345–356.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1169–1178.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125*(1), 3–30.

Russell. J. A., Lewicka, M., & Nitt, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology, 57*(5), 848–856.

Russell, J. A., & Ridgeway, D. (1983). Dimensions underlying children's emotion concepts. *Developmental Psychology, 19*(6), 795–804.

Soler, C., Núñez, M., Gutiérrez, R., Núñez, J., Medina, P., Sancho, M., Álvarez, J., & Núñez, A. (2003). Facial attractiveness in men provides clues to semen quality. *Evolution and Human Behavior, 24*(3), 199–207.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human–Computer Interaction, 11*(2), 97–124.

Subramony, D. (2002). Introducing a "means-end" approach to human–computer interaction: Why users choose particular web sites over others. In P. Barker & S. Rebelsky (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 1886–1891). Chesapeake, VA: AACE.

Suchman, L. A. (1987). *Plans and situated actions: The problem of human–machine communication.* New York, NY: Cambridge University Press.

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences*, *106*(43), 18362–18366.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human–robot interaction. *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction* (pp. 33–40). March 2–3, Salt Lake City, USA.

Sung, J.-Y., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "My Roomba is Rambo": Intimate home appliances. *In Proceedings of the 9th international conference on Ubiquitous computing* (Ubicomp '07) (pp. 145–162). September 16–19, Innsbruck, Austria.

Tinwell, A., & Grimshaw, M. (2009). Bridging the uncanny: An impossible traverse? *Proceedings of the 13th International Mindtrek Conference: Everyday Life in the Ubiquitous Era* (pp. 66–73). September 30–October 2, Tampere, Finland.

Turkle, S. (2007). The things that matter. In S. Turkle (Ed.), *Evocative objects: Things we think with* (pp. 3–10). Cambridge, MA: MIT Press.

Turkle, S., Taggart, W., Kidd, C. D., & Daste, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. *Connection Science, 18*(3), 347–361.

Uekermann, F., Herrmann, A., Wentzel, D., & Landwehr, J. R. (2008). The influence of stimulus ambiguity on category and attitude formation. *Review of Managerial Science, 4*(1), 33–52.

Valette-Florence, P. (1998). A causal analysis of means-end hierarchies in a cross-cultural context: Methodological refinements. *Journal of Business Research, 42*(2), 161–166.

Van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts and what model to use instead. *Applied Psychological Measurement, 18*(2), 97–110.

Vanden Abeele, P. (1992). A means-end study of dairy consumption motivation. *Report for the European Commission*, EC Regulation 1000/90–43 ST.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*(2)*,* 219–235.

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations and psychobiological evidence. *Journal of Personality and Social Psychology, 76*(5), 820–838.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Science, 14*(8), 383–388.

Whorf, B. L. (1940). Science and Linguistics. *Technology Review, 42*(6), 229–248.

Wojciszke, B., Abele, A. E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology, 39*(6), 973–990.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1245–1257.

Woods, S., Dautenhahn, K., & Schulz, J. (2005). Child and adults' perspectives on robot appearance. *Proceedings of the Symposium on Robot Companions* (pp. 126–135), April 12–15, Hatfield, England.

Yamauchi, T. (2005). Labeling bias and categorical induction: Generative aspects of category information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 538–553.

Zaman, B., & Abeele, V. V. (2010). Laddering with young children in user experience evaluations: Theoretical groundings and a practical case. *Proceedings of the 9th International Conference on Interaction Design and Children* (IDC '10) (pp. 156–165), New York, NY.

Zimmerman, D. E., & Akerelrea, C. (2002). A group card sorting methodology for developing informational web sites. *Proceedings of International Professional Communication Conference* (pp. 437–445), September 17–20, Portland, OR.

CURRICULUM VITAE

Chin-Chang Ho

Education

Ph.D.	Informatics, Indiana University, Indianapolis, IN, USA	2015
M.S.	Human-Computer Interaction, Indiana University, Indianapolis, IN, USA	2008
M.S.	Social Informatics, Yuan-Ze University, Taiwan	2002
B.S.	Clinical Psychology, Fu-Jen Catholic University, Taiwan	1999

Work Experiences

Research Assistant

Graduate School of Informatics, IUPUI — 2007.01~2011.05

Teaching Assistant

"I563-Psychology of Human-Computer Interaction", Graduate Course, School of Informatics, IUPUI — 2008.01~2008.05

Teaching Assistant

"I575-Informatics Research Design" Graduate Course, School of Informatics, IUPUI — 2007.01~2007.05

Research Assistant

National Center for High-performance Computing, Taiwan 2004.08~2005.07

"The Mechanism to Sustain Scientific Volunteers"

Contributing Translator

2002.09~2003.02

PC Magazine Republic of China Edition

Research Assistant

National Science Council, Taiwan

2001.09~2002.07

"The Study of Unemployment and Inadequate Employment in the Information Society."(NSC90-2412-H-155-002-SSS)

Teaching Assistant

"Data Collection and Analysis" Undergraduate Course, 2001.03~2001.06

Department of Information Communication, Yuan-Ze University

Teaching Assistant

"Research Method" Undergraduate Course, Department of 2000.09~2001.01

Information Communication, Yuan-Ze University

Research Assistant

National Science Council, Taiwan

2000.09~2001.07

"The Study of Information Gap and The Mobility of Information Class in Taiwan," (NSC89-2412-H-155-003-SSS)

Teaching Assistant

"Media and Society" Undergraduate Course, Department of 2000.03~2000.06

Information Communication, Yuan-Ze University

Teaching Assistant

"Communication Theories" Undergraduate Course, Department of Information communication, Yuan-Ze University 1999.09~2000.01

Research Assistant

Department of Health, Taiwan

"Information Ethic and Data Security in Medical Data Bank" (DOH89-TD-1090) 1999.09~2000.07

Internship

Psychiatry Department, Taipei Municipal Yang Ming Hospital, Taipei, Taiwan 1999.02~1999.06

Internship

Psychiatry Department, Shin-Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan 1998.09~1999.01

Journal Articles

Mitchell, W. J., Ho, C.-C., Patel, H., & MacDorman, K. F. (2011). Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior, 27*(1), 402–412.

Faiola, A., Ho, C.-C., Tarrant, M. A., & MacDorman, K. F. (2011). The aesthetic dimensions of US and South Korean responses to web home pages: A cross-cultural comparison.*International Journal of Human-Computer Interaction, 27*(2), 131–150.

MacDorman, K. F., Whalen, T. J., Ho, C.-C., & Patel, H. (2011). An improved scale for measuring usability from novice and expert performance. *International Journal of Human-Computer Interaction, 27*(3), 1–23.

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior, 26*(6), 1508–1518.

MacDorman, K. F., Coram, J. A., Ho, C.-C., & Patel, H. (2010). Gender differences in the impact of presentational factors in human character animation on decisions in ethical dilemmas. *Presence: Teleoperators and Virtual Environments, 19*(3), 213–229.

MacDorman, K. F., Vasudevan, S. K., & Ho, C.-C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society, 23*(4), 485–510.

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. (2009). Too real for comfort: Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695–710.

Green, R. D., MacDorman, K. F., Ho. C.-C. & Vasudevan, S. K. (2008). Sensitivity to Proportions in Faces of Varying Human Likeness. *Computers in Human Behavior. 24*(5), 2456–2474.

MacDorman, K. F., Ough, S., & Ho, C.-C. (2007). Automatic emotion prediction of song excerpts: Index construction and algorithm design and empirical comparison. *Journal of New Music Research,* 36(4), 283–301.

Ho, C. & Tseng, S. (2006). From Digital Divide to Digital Inequality-The Global Perspective. *International Journal of Internet and Enterprise Management*. *4*(3), 215–227.

Tseng, S., You, Y. & Ho, C. (2002). New Economy, Underemployment and Inadequate Employment. *Journal of Cyber Culture and Information Society*, *3*, 215–237.

Tseng, S. Hsieh, Y. & Ho, C. (2001). The Computerization and Protection of Electronic Patient Records in Hospitals. *The Journal of Taiwan Association for Medical Informatics*, *13*, 19–42.

Hsieh, Y. & Ho, C. (2001). The Development of Nation State in Information Society. *Journal of Cyber Culture and Information Society*, 1, 201–228.

Conference Paper

Ho, C.-C., MacDorman, K. F., & Pramono, Z. A. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings. *Proceedings of the Third ACM/IEEE International Conference on Human-Robot Interaction* (pp. 169–176). March 11–14. Amsterdam.

Ho, C.-C., Tseng, S.-F., & Huang, H.-I. (2005). Academic productivity, coordinated problem and cultural conflict in the scientific collaboration community. *Proceedings of the Annual Meeting of American Sociological Association,* Philadelphia, PA.

Ho, C.-C., & Tseng, S.-F. (2003). From digital divide to digital inequality: The global perspective. *Proceedings of the Annual Meeting of American Sociological Association,* Atlanta, GA.

You, Y.-C., & Ho, C.-C. (2002). Inadequate employment and mismatch: The underemployment in Taiwan. *Proceedings of the Annual Meeting of American Sociological Association,* Chicago, IL.

Tseng, S.-F., You, Y.-C., & Ho, C.-C. (2002). New economy, underemployment, and inadequate employment. *Proceedings of the 3rd International Congress of the Work & Labour Network: Labour, Globalisation and the New Economy,* Osnabruck, Germany.

Tseng, S.-F., & Ho, C.-C. (2001). The global digital divide and social inequality: Universal or polarized? *Proceedings of the Annual Meeting of American Sociological Association,* Anaheim, CA.

Tseng, S.-F., & Ho, C.-C. (2001). The usage and evaluation of e-Health provider website. *Proceedings of the Annual Meeting of Chinese Communication Society*, Hong Kong.

Tseng, S.-F., & Ho, C.-C. (2000). The privacy concerns and ethics in maintaining electronic patient records. *Proceedings of the Annual Meeting of American Sociological Association,* Washington, DC.

Posters

Srinivas, P., Patel, H., Ho, C.-C., & MacDorman, K. F. (2011). An uncanny valley of visual perspective taking: A study of the effete of character human likeness and eeriness on altercentric intrusions during a dot counting task. IUPUI Research Day. April 8, 2011. Indianapolis, Indiana.

MacDorman, K. F., Ho, C.-C., Lu, Amy S., Mitchell, W. J., Patel, H., & Srinivas, P. (2011). Decision making, empathy, and the uncanny valley: Our inferences about virtual humans depend on their appearance, speech, and motion quality. IUPUI Research Day. April 8, 2011. Indianapolis, Indiana.

MacDorman, K. F., Gadde, P., Ho, C.-C., Mitchell, W. J., Patel, H., Schermerhorn, P. W., & Scheutz, M. (2010). Probing people's attitudes and behaviors using humanlike agents. IUPUI Research Day. April 9, 2010. Indianapolis, Indiana.

Master Thesis

Human Emotion and the Uncanny Valley: A GLM, MDS, and ISOMAP Analysis of Robot Video Ratings. Indiana University-Purdue University, Indianapolis.

The Implementation of Medical Information Technology and the Changing Medical Professions. Yuan-Ze University, Taiwan.

Research Interests

Human-Robot Interaction

Human-Computer Interface

The Social Inequality between Technological and Economic Development

Quantitative Research Method and Statistics

Computer Skills

Languages: HTML, PHP, XML, CSS

Databases: MySQL, MS SQL

Statistics: SPSS, SAS, R, STATA, LISERL, AMOS, UCINET, Pajek, HLM.

Miscellaneous: MATLAB, Axure, Photoshop, Maya, ZBrush