Int J of Soc Robotics manuscript No.
(will be inserted by the editor)

Measuring the Uncanny Valley Effect

Refinements to Indices for Perceived Humanness, Attractiveness, and Eeriness

Chin-Chang Ho · **Karl F. MacDorman**

Received: date / Accepted: date

Abstract Using a hypothetical graph, Masahiro Mori proposed in 1970 the relation between the human likeness of robots and other anthropomorphic characters and an observer's affective or emotional appraisal of them. The relation is positive apart from a *U*-shaped region known as the *uncanny valley*. To measure the relation, we previously developed and validated indices for the perceptual-cognitive dimension *humanness* and three affective dimensions: interpersonal *warmth, attractiveness,* and *eeriness*. Nevertheless, the design of these indices was not informed by how the untrained observer perceives anthropomorphic characters categorically. As a result, scatter plots of humanness vs. eeriness show the stimuli cluster tightly into categories that are widely separated from each other. The present study applies a card sorting task, laddering interview, and adjective evaluation ($N = 30$) to revise the humanness, attractiveness, and eeriness indices and validate them via a representative survey ($N =$1,311). The revised eeriness index maintains its orthogonality to humanness ($r = .04$, $p = .285$), but the stimuli show much greater spread, reflecting the breadth of their range in human likeness and eeriness. The revised indices enable empirical relations among characters to be plotted similarly to Mori's graph of the uncanny valley. Accurate measurement with these indices can be used to enhance the design of androids and 3D computer-animated characters.

Keywords Anthropomorphism · Categorical perception · Cognitive bias · Embodied agents · Human realism · Psychometric scales · Social perception

Chin-Chang Ho · Karl F. MacDorman
Indiana University School of Informatics and Computing, 535 West Michigan Street, Indianapolis, IN 46202 USA
E-mail: kmacdorm@indiana.edu

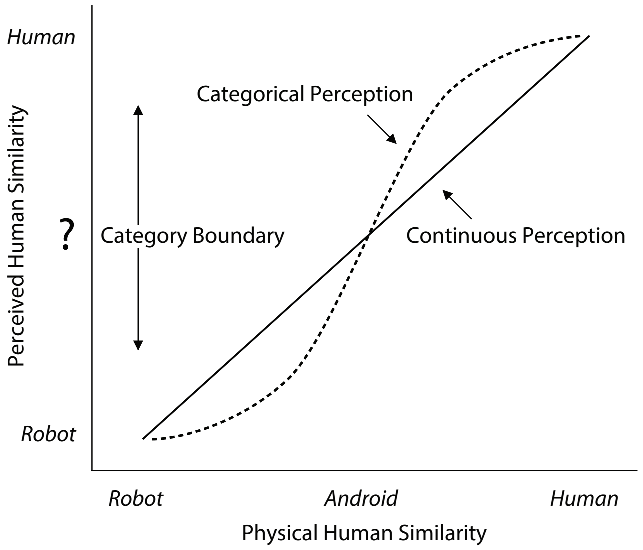

Fig. 1 Categorical perception causes equal-sized differences in a character's physical similarity to a human being to appear much smaller within the category *robot* or *human* than at the boundary between them.

1 Introduction

Human physical and behavioral realism heightens empathy for robots, which in turn enhances social interaction [34, 51]. This is advantageous in settings where it is preferred to have the observer expect humanlike performance from the robot, such as a robot portraying a standardized patient during a trainee's assessment [18,21]. It is also helpful to measure how the observer perceives human realism and evaluates it affectively to develop design principles for increasing human acceptance of android robots and three-dimensional (3D) computer-animated characters.

Accurate measurement is vital because humanlike characters are susceptible to negative affective evaluations known as the uncanny valley effect [35,33,42]. These evaluations have been characterized as cold, eerie feelings, associated with, but not equivalent to, fear, anxiety, and disgust, a loss of empathy, and avoidance behavior [23,31,37, 41]. Mori [44] illustrated the uncanny valley effect by drawing a valley of eeriness in a graph that otherwise depicts a positive relation between human likeness and affinity. A similar curve approximated ratings of a large sample of real-world robots, though with considerable variance ($R^2_{\text{adj}} = .29$) [38]. However, there has been insufficient research on how to measure the uncanny valley effect accurately.

Various methods have been used to evaluate human–robot interactions, including spatial engagement, open-ended questions, and principal component analysis [40,45,46,61, 62]. In the context of the uncanny valley, the present study focuses specifically on scale development for evaluating android robots and 3D computer-animated characters.

Bartneck [2] proposed the Godspeed indices, which were designed to measure anthropomorphism, animacy, likability, perceived intelligence, and perceived safety. These indices average ratings on semantic differential scales. Unfortunately, the anthropomorphism, animacy, and likability indices were highly correlated ($.67 < r < .89$, $p < .001$) [22] and thus may not measure distinct concepts. This is a reoccurring measurement issue, because humanness tends to be associated with other positive social attributes.

Although Ho and MacDorman's [22] perceived humanness, attractiveness, and eeriness indices have high internal reliability and the correlation of eeriness with humanness, attractiveness, and warmth was not significant, their scatter plots form two widely separated clusters: (a) mobile, humanoid, and android robots and (b) 3D computer models of humans that range from the cartoon-like to the photorealistic. Each cluster is tightly grouped despite the varied appearance of the characters within it (Figure 4). The formation of two tightly grouped but widely separated clusters indicates the presence of categorical perception effects in observing the characters' appearance or behavior. This pattern occurs if one bipolar adjective describes one perceptual category and the other bipolar adjective describes a different perceptual category, because category perception causes physical difference among stimuli within each category to appear much smaller than equal-sized differences between categories (e.g., robot vs. human, Figure 1) [13,20,43].

These indices should be designed to span category boundaries because anthropomorphic entities whose features span them are prone to elicit the uncanny valley effect [7,30,39,42,43,54]. These negative evaluations are likely to persist at least until a new category is formed and labeled [58]. As stimuli that span the new category and its neighbors are categorized, categorical perception then develops along those continua [6,25].

Categorical perception, also called the perceptual magnet effect, has recently been found on the continuum from 3D computer models to photographs of real people [7–9,24, 27,30,35]. Various theories have been proposed that broadly relate categorization to the uncanny valley, including theories that are based on categorical perception [50], categorization difficulty [63], cognitive dissonance [19], balance theory [56], and feature inconsistency [42,43]. The categorical perception of humanlike characters necessitates examining how observers categorize the characters to ensure that the humanness, attractiveness, and eeriness indices adequately represent within-category variation along these three dimensions and span category boundaries.

The present study seeks to improve these indices for measuring the uncanny valley effect in light of how observers' categorize mobile, humanoid, and android robots and 3D computer-animated characters. To address the effects of categorical perception and anthropomorphism, card sorting is applied to determine how untrained observers categorize robots, computer-animated characters, and real human beings, thus revealing their own categories and the boundary regions between them. The bipolar adjectives of the semantic differential scales composing these indices are next evaluated to determine adjective pairs that span the categories and their boundary regions. The resulting indices are then evaluated in a representative survey. Improving measurement instruments for the uncanny valley is significant both scientifically, in more accurately describing the phenomenon and evaluating its effects, and in testing proposed design principles for overcoming the uncanny valley.

2 Method

This study applied a four-stage exploratory sequential design that sought to improve the humanness, attractiveness, and eeriness indices [22]: (1) a card sorting task to probe how each participant conceptualizes humanlike characters; (2) a laddering interview to collect new candidate adjectives to revise the semantic differential scales that comprise the indices; (3) a bipolar adjective evaluation to verify the importance of the original scales to the categories identified by the participant; and (4) a representative web survey to validate the revised indices. This study was approved by the Indiana University Institutional Review Board (EX0903-35B).

2.1 Participants

For the card sorting task, laddering interview, and adjective evaluation, 30 participants were recruited by email and flyers in a convenience sample from a Midwestern U.S. public university system: 70.0% were male, 30.0% female, and the median age was 26. Participants completed these stages from January to June 2013. There was no attrition. Participants received a $10 gift card.

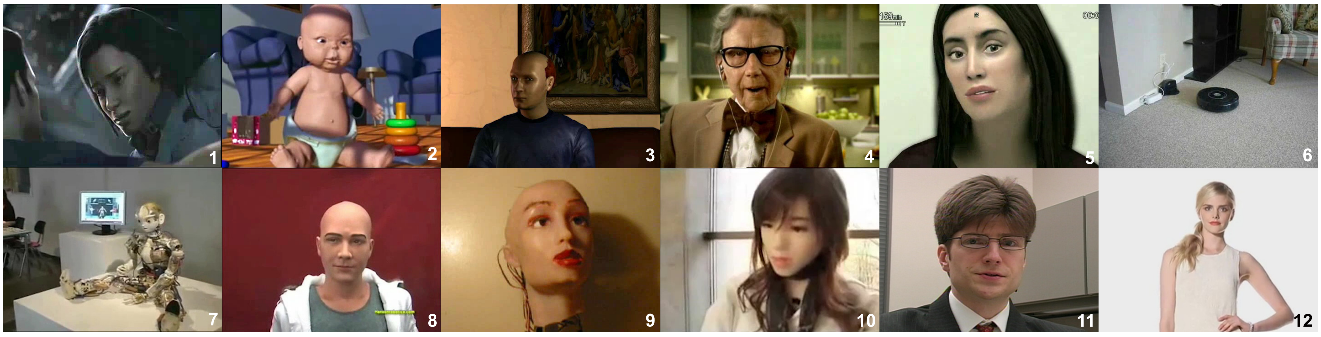

Fig. 2 The 12 characters are five 3D computer animations, (1) Doctor Aki Ross from the film *Final Fantasy: The Spirits Within* (2001), (2) Billy, the baby from “Tin Toy” (1988), (3) an unnamed man from Phil Rice’s “Apology” (2008), (4) Orville Redenbacher from a popcorn commercial (2007), and (5) Mary Smith from “Heavy Rain: The Casting” (2006), five robots, (6) Roomba 570 (iRobot), (7) Kotaro (JSK, University of Tokyo), (8) Jules (Hanson Robotics), (9) Animatronic Head (David Ng), and (10) Aiko (Le Trung), and two human beings, (11) a man and (12) a woman.

For the web survey, 1,311 participants were recruited by email in a random sample of undergraduate students from the same university system: 39.1% were male, 60.9% female, 81.5% were age 18–25, 5.4% 26–30, and 13.1% 31 or older. The sample population was 74.1% white, 7.3% African American, 5.5% Hispanic/Latino, 3.1% Asian, 0.2% American Indian, 0.1% Pacific Islander, 2.9% two or more races, 5.8% international, and 1.1% unknown. Additional inclusion criteria were 18 or older, native English speaker, and 20/40 vision or better with correction. Some data was missing at random owing to attrition. The measurement error range was ±2.89% at the 95% confidence level. Participants completed the web survey from March to April 2014. Participants received no compensation.

2.2 Materials and Procedures

In the card sorting task, each of the 30 participants viewed a randomized sequence of video clips that corresponded to 12 different characters: five 3D computer-animated characters, five robots, and two human beings (Figure 2). One of the robots (Hanson Robotics’ Elvis) and two of the 3D computer-animated characters (from *The Incredibles* and *The Polar Express*) from Ho and MacDorman (2010) were replaced to improve representativeness. The aim was to select robots from typical demonstration settings and 3D computer models from a variety of genres—short films, machinima,[1] advertisements, and videogames—in addition to feature-length films. Two humans were added to extend the range of humanness. The video clips were 480 pixels by 360 pixels (a 4:3 aspect ratio) and were 15 to 30 seconds in duration.

A representative frame from each video was printed in color on a $3\frac{1}{2}$-by-5-inch card. Using the cards as visual aids, the participant grouped the 12 characters into self-determined categories and proposed a label for each category [53]. The participant was instructed to sort each character into only one category, thus ensuring that the categories were mutually exclusive. The experimenter used prompts, such as “Which characters would you group together, or separate from the others?” The participant then verified the categories by reviewing the video clips at least once.

Next, in the laddering interview, the participant was asked to list the characteristics of each character. For each characteristic, the participant was asked repeatedly, “Why is that important to you?” The participant’s answer typically linked a formal characteristic like “mechanical movement” to an aesthetic judgment like “mismatched with human appearance” to an experiential characteristic like “weird.” The participant was required to provide at least three laddering responses.

Finally, in the adjective evaluation, the participant rated on a 3-point importance scale (1. *slightly important,* 2. *moderately important,* 3. *very important*) all bipolar adjectives comprising the humanness (12 adjectives), attractiveness (10 adjectives), and eeriness (16 adjectives) indices for each category that the participant had proposed in the card sorting task [22,55,60]. Each semantic differential scale is comprised of a bipolar adjective that is low on the scale (e.g., *artificial* is low on the humanness scale) and a bipolar adjective that is high on the scale (e.g., *natural* is high on the same scale). If the participant considered both bipolar adjectives important (e.g., *artificial* and *natural*), the scale was expected to measure the concept effectively; if the participant only considered one of the bipolar adjectives important, the adjectives might not span that category.

In the representative web survey, each participant rated the 12 characters on the semantic differential scales comprising the three indices, while the corresponding video clip played in a continuous loop. The semantic differential scales included new candidate adjectives from the laddering interview. As before, the characters were presented one at a time and in random order. Scale order was also randomized. The

[1] The cinematic production of narrative computer animation by means of a videogame or other real-time graphics engine.

semantic differential scales recorded a 7-point value ranging from -3 to $+3$.

2.3 Data Analysis Procedures

A semantic differential scale is defined as unbalanced to the extent that one bipolar adjective is important for more categories than the other. In taking the difference in matches between the low and high adjective of a scale, the magnitude represents the degree of imbalance and the sign represents the direction. For example, if a participant proposed three categories — Animation, Humanlike Robot, and Real Human — and stated that *natural* was important for all three categories but *artificial* only for Humanlike Robot, the *natural–artificial* scale is unbalanced ($3 - 1 = +2$). Based on the bipolar adjective evaluation, if the mean imbalance of a scale was statistically significant, an alternative bipolar adjective was tested.

For the web survey, three criteria for bipolar adjective selection were applied: (a) high internal reliability, (b) loading on the correct factor, and (c) correlation with the ‘sanity check’ scale. Internal reliability of the indices was assessed with Cronbach’s α. To determine whether the semantic differential scales loaded on factors matching their named concepts, exploratory factor analysis was used, namely, principal component analysis with Promax rotation [17].

To verify that each index measured its concept, the following sanity check scales were included: *artificial–natural, unattractive–attractive,* and *reassuring–eerie* for the humanness, attractiveness, and eeriness index, respectively. Sanity check scales have face validity but do not meet one or two other criteria. If a scale of a particular index did not load on the same dimension as its sanity check scale or if its factor loading was low ($< .40$), the scale was removed from the index. The sanity check scales were excluded from the final set of revised indices.

A correlation analysis was used to evaluate the discriminant validity of the indices and the degree to which humanness was decorrelated from attractiveness and eeriness. Confirmatory factor analysis further verified the construct validity of the revised indices. Significance in comparing groups was assessed by a one-way analysis of variance (ANOVA). To visualize relations among the semantic differential scales of the indices, multidimensional scaling (MDS) was employed to reduce 18 dimensions to 2.

SPSS (ver. 20) was used to perform internal reliability assessment, exploratory factor analysis, and correlation analysis, LISREL (ver. 8.54) to perform confirmatory factor analysis, and MATLAB (ver. 8.5) to perform multidimensional scaling.

Cronbach’s α thresholds were interpreted as acceptable $= .7$, good $= .8$, and excellent $= .9$. The factor loading cutoff for scale removal was .40 for exploratory factor analysis and .60 for confirmatory factor analysis. Test statistics were interpreted with a significance threshold of $\alpha = .05$.

3 Results

3.1 Card Sorting

All 30 participants proposed to group the 12 characters in at least two categories. More than half (54%) proposed at least four categories ($M = 4.38$), thus exceeding the three nominal categories of robots, animations, and humans. The final categories (in decreasing order of frequency) were human ($n = 16$), robot (15), animation (14), machine (5), android (3), man (3), woman (3), 3D character (2), advance robot (2), advertisement (2), cartoon (2), digital creation (2), dummy (2), half human–half robot (2), humanlike robot (2), Japanese doll (2), machine part (2), prototype (2), robot machine (2), and utility robot (2). For the anthropomorphic characters, participants often preferred to use narrower categories (e.g., *advanced robot*) instead of broader ones (e.g., *robot*). Even though the participants identified various categories, only three used android specifically.

3.2 Scale Evaluation

Of the 38 bipolar adjectives evaluated with respect to the perceived categories, those comprising the semantic differential scales of the humanness index were deemed most important ($M = 2.00$, $SD = 0.25$, $n = 30$), followed by attractiveness ($M = 1.64$, $SD = 0.40$) and eeriness ($M = 1.60$, $SD = 0.33$). However, when categorizing the anthropomorphic characters, the participants were more likely to choose low humanness adjectives ($M = -0.34$, $SD = 1.24$), low eeriness adjectives ($M = -0.24$, $SD = 0.63$), and high attractiveness adjectives ($M = 0.33$, $SD = 0.82$).

Adjective importance was compared for robot-related categories versus the other categories and likewise for animation-related and human-related categories versus the other categories using a one-way ANOVA (Table 1). Fewer humanness adjectives were used for animation-related categories ($M = 1.87$, $SE = 0.07$) than for other categories ($M = 2.03$, $SE = 0.03$, $F(1,61) = 4.37$, $p = .041$) and more attractiveness adjectives were used for human-related categories ($M = 1.85$, $SE = 0.12$) than for other categories ($M = 1.57$, $SE = 0.05$, $F(1,61) = 6.18$, $p = .016$).

Imbalance in the importance of bipolar adjectives was similarly compared. More low humanness adjectives (e.g., *inanimate*) were used for robot-related categories ($M = -1.12$, $SE = 0.16$) than for other categories ($M = 0.11$, $SE = 0.20$, $F(1,61) = 18.57$, $p < .001$) and more high humanness adjectives for human-related categories ($M = 1.56$, $M = 0.11$) than for other categories ($M = -0.94$, $SE = 0.10$,

Table 1 Adjective importance and imbalance in bipolar adjective importance by category

	Adjective Importance			Importance Imbalance		
	Humanness	Eeriness	Attractiveness	Humanness	Eeriness	Attractiveness
Robot-related	2.07	$1.70^{\triangle}$	1.62	$-1.12^{\ddagger}$	−0.15	0.22
Others	1.96	1.55	1.66	0.11	−0.29	0.40
Animation-related	1.87^{*}	1.53	1.52	−0.82	−0.12	0.15
Others	2.03	1.62	1.67	−0.21	−0.27	0.38
Human-related	2.09	1.62	1.85^{*}	$1.56^{\ddagger}$	$-0.70^{\ddagger}$	$0.83^{\dagger}$
Others	1.97	1.60	1.57	−0.93	−0.09	0.18

$^{\triangle}p < .1.$ $^{*}p < .05.$ $^{\dagger}p < .01.$ $^{\ddagger}p < .001.$

$F(1, 61) = 172.93$, $p < .001$). More low eeriness adjectives were also used for human-related categories ($M = -0.70$, $SE = 0.13$) than for other categories ($M = -0.09$, $SE = 0.09$, $F(1, 61) = 12.47$, $p < .001$) and more high attractiveness adjectives for human-related categories ($M = 0.83$, $SE = 0.20$) than for other categories ($M = 0.18$, $SE = 0.11$, $F(1, 61) = 7.91$, $p = .007$).

3.3 Revised Scales

Bipolar adjectives differed in their rated importance depending on the category. For each category, the difference in importance between the low and high bipolar adjective of each semantic differential scale was compared to identify imbalance in their relative importance.

The results indicate that for the robot category, the scale *without definite lifespan–mortal* ($p = .006$) of the humanness index was significantly unbalanced, as were the scales *numbing–freaky* ($p = .005$) and *unemotional–hair-raising* ($p = .002$) of the eeriness index, thus indicating these scales required revision. For the animation category, two scales of the humanness index were significantly unbalanced: *synthetic–real* ($p = .007$) and *mechanical movement–biological movement* ($p = .014$).

For the human category, *inanimate–living* ($p = .001$) of the humanness index was significantly unbalanced. Three scales of the eeriness index were significantly unbalanced: *reassuring–eerie* ($p = .007$), *ordinary–supernatural* ($p < .001$), and *unemotional–hair-raising* ($p = .019$). Two scales of the attractiveness index were significantly unbalanced: *unattractive–attractive* ($p = .034$) and *crude–stylish* ($p = .013$).

For the android category, two scales of the eeriness index were significantly unbalanced: *numbing–freaky* ($p = .014$) and *unemotional–hair-raising* ($p = .029$). Three scales of the eeriness index were unbalanced: *numbing–freaky, ordinary–supernatural,* and *unemotional–hair-raising*.

Using the laddering responses as a pool of candidate adjectives, we tentatively considered *dull–freaky* and *boring–freaky* as potential replacements for *numbing–freaky; ordinary–unreal* and *ordinary–creepy* for *ordinary–supernatural*; *unemotional–alarming* for *unemotional–hair-raising*; and *predictable–eerie* for *reassuring–eerie*. In addition, *plain–weird, conformist–bizarre,* and *habitual–supernatural* were also considered. These new scales were then included in the web survey with the original ones to test whether they were more appropriate for untrained observers.

3.4 Validation of New Scales

The five scales of the humanness index were validated: *inanimate–living, synthetic–real, mechanical movement–biological movement, human-made–human-like,* and *without definite lifespan–mortal,* and the sanity check *artificial–natural*. Overall internal reliability was good (Cronbach's $\alpha = .84$).[2] The exploratory factor analysis showed all five scales and the sanity check, loaded on one factor, which explained 58.30% of the total variance. These results confirmed the reliability and validity of the original humanness index [22].

The four scales of the attractiveness index were validated: *ugly–beautiful, crude–stylish, repulsive–agreeable,* and *messy–sleek,* and the sanity check *unattractive–attractive*. Overall internal reliability of the index was good (Cronbach's $\alpha = .88$). Exploratory factor analysis showed all four scales, including the sanity check, loaded on a single factor that explained 65.08% of the total variance. These results confirmed with a new sample the reliability and validity of the original attractiveness index [22].

All seven scales comprising the original eeriness index and its sanity check were validated. Factor analysis confirmed the existence of the two subfactors of the eeriness index previously found in Ho and MacDorman [22]. *Uninspiring–spine-tingling, boring–shocking, predictable–thrilling, bland–uncanny,* and *unemotional–hair-raising* loaded on the spine-tingling subfactor, which explained 39.54% of the total variance with a Cronbach's α of .84. *Reassuring–eerie, numbing–freaky,* and *ordinary–supernatural* loaded on the eerie subfactor, which explained 23.62% of the total vari-

[2] The value is the mean of 12 Cronbach's αs, one for each character.

Table 2 Factor loadings of the revised semantic differential scales

	Humanness	Eeriness			Attractiveness
		Eerie		Spine-tingling	
Inanimate–Living	.81				
Synthetic–Real	.80				
Mechanical Movement–Biological Movement	.77				
Human-Made–Humanlike	.76				
Without Definite Lifespan–Mortal	.67				
Dull–Freaky$^\diamond$		.76			
Predictable–Eerie$^\diamond$		.75			
Plain–Weird$^\diamond$		.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	.86	.81	.85

Model fit: $\chi^2 = 3783$, $df = 129$, GFI = .95, AGFI = .93, NFI = .97, CFI = .97, RMR = .15, RMSEA = .061
$^\diamond$New candidate scale

ance. However, the Cronbach's α of the eerie subfactor was only .69, indicating the need to improve its reliability.

Seven candidate scales, *dull–freaky, ordinary–unreal, ordinary–creepy, plain–weird, predictable–eerie, conformist–bizarre,* and *habitual–supernatural* loaded on the same dimension as *reassuring–eerie, numbing–freaky,* and *ordinary–supernatural;* two scale candidates, *unemotional–alarming* and *boring–freaky,* loaded on the same dimension as *boring–shocking, uninspiring–spine-tingling, predictable–thrilling, bland–uncanny,* and *unemotional–hair-raising.*

First, the candidates *ordinary–creepy* ($r = .70$) and *habitual–supernatural* ($r = .71$) highly correlated with the dimension of the original set, *reassuring–eerie, numbing–freaky,* and *ordinary–supernatural,* indicating these scales were redundant and thus should be excluded. Second, adding the candidates *unemotional–alarming* and *boring–freaky* only slightly increased the internal reliability of the spine-tingling subfactor (Cronbach's αs ranged from .84 to .86), indicating this subfactor, which included *uninspiring–spine-tingling, boring–shocking, predictable–thrilling, bland–uncanny,* and *unemotional–hair-raising,* was already saturated. Given that these five reliable scales were already available to measure the concept, we did not need to develop any additional scales. Therefore, *unemotional–alarming* and *boring–freaky* were excluded from the revised index. Third, *ordinary–creepy* ($r_{\text{attr}} = -.45$, $r_{\text{hum}} = -.31$), *ordinary–unreal* ($r_{\text{attr}} = -.37$, $r_{\text{hum}} = -.44$), *conformist–bizarre* ($r_{\text{attr}} = -.35$, $r_{\text{hum}} = -.28$), and *numbing–freaky* ($r_{\text{attr}} = -.30$, $r_{\text{hum}} = -.23$) significantly correlated with the attractiveness and humanness indices, which violated the criterion of scale decorrelation (cf. [22]). Therefore, they were excluded from the revised index. (*Ordinary–supernatural* was retained, despite its bias, because the alternative candidates, *ordinary–unreal* and *ordinary–creepy*, loaded on both the eerie and spine-tingling subfactors.)

Based on the three criteria for bipolar adjective selection (i.e., high internal reliability, loading on the correct factor, and correlation with the sanity check scale), four scales were developed for an revised version of the attractiveness index, nine scales for the eeriness index, and five scales for the humanness index.

Confirmatory factor analysis was employed to verify the theoretical structure of this final set of 18 semantic differential scales (shown in Table 2 with their factor loadings). Although one index (RMSEA = .061) exceeded the cutoff of .05, the remaining indices indicated the 18 semantic differential scales had high goodness-of-fit within the structure of the humanness, eerie, spine-tingling, and attractiveness indices ($\chi^2 = 3783$, CFI = .97, NFI = .97, GFI = .95, AGFI = .93) [5,10,16]. The revised scales showed improved fit as compared with those of Ho and MacDorman [22] (RMSEA decreased from .075 to .061, GFI increased from .91 to .95, and AGFI increased from .88 to .93). Further, the statistics of goodness-of-fit indicated the eerie and spine-tingling subfactors of the eeriness index were robust enough to represent their own theoretical construct ($r = .44$).

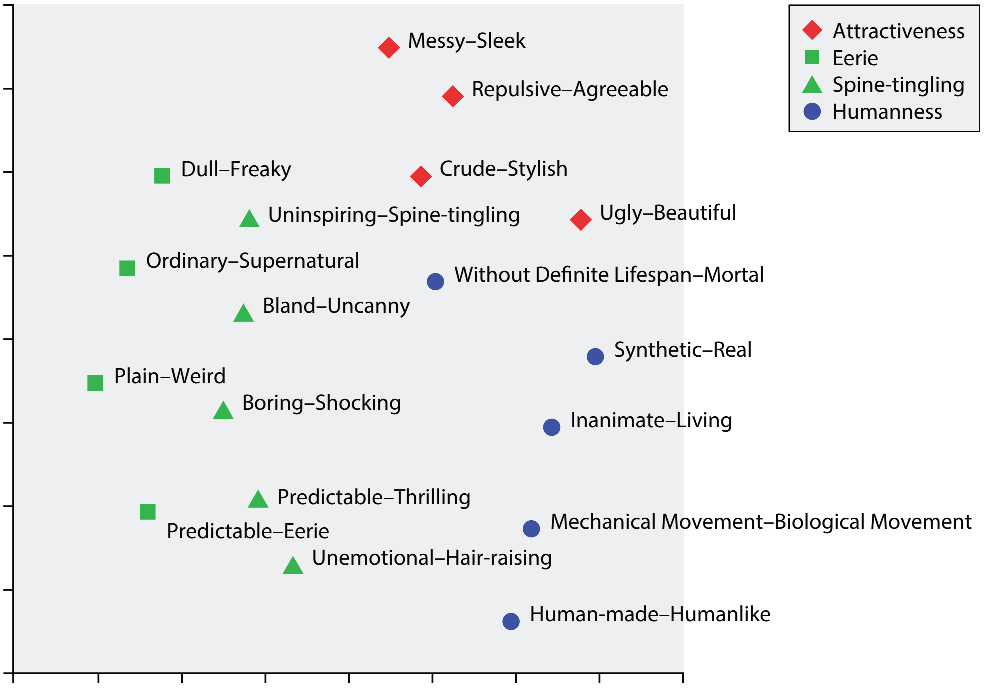

Fig. 3 Multidimensional scaling was performed on the 18 semantic differential scales using the ratings of the characters in the 12 video clips. The scales of the humanness, eerie, spine-tingling, and attractiveness indices were well separated.

The correlation analysis indicated the revised indices retained their construct validity (Table 3). Eeriness had no significant correlation with either humanness or attractiveness, reflecting its good discriminant validity.

Multidimensional scaling was performed on the 18 semantic differential scales of the humanness, attractiveness, and eeriness indices. The scales occupied three well separated, nonoverlapping regions (Figure 3). Furthermore, for the eeriness index, the four scales of its eerie subfactor and the five scales of its spine-tingling subfactor occupied two well separated, nonoverlapping regions. The MDS results show the humanness, attractiveness, and eeriness indices distinctly measured their concepts.

In comparing the scatter plot of stimuli from Ho and MacDorman [22] (Figure 4) with those from this study (Figure 5), the revised humanness and eeriness indices better capture the extent of within-category variation, thus mitigating the effects of categorical perception. The internal reliability of the eeriness index also increased from acceptable (Cronbach's $\alpha = .74$) to good (.86).

Table 3 Correlation between the revised humanness, attractiveness, and eeriness indices

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

4 Discussion

The categorization task revealed how observers apply categories in perceiving humanlike characters [36]. The categories supported inferences both about attributes of the character and about unrelated attributes [64]. Although the study's untrained participants placed greater importance on the humanness bipolar adjectives than the attractiveness and eeriness ones, they tended to use the latter more frequently when evaluating the characters.

An evaluation of the scales comprising the humanness, attractiveness, and eeriness indices with respect to self-identified categories revealed that some pairs of bipolar adjectives were unbalanced in their importance. If one pole of a scale is unimportant for all sorted characters in a given category, that scale is unlikely to measure differences effectively within the category along the corresponding dimension.

During the card sorting task, untrained participants found it challenging to partition humanlike characters on a humanness continuum. Instead, they relied on their prior domain knowledge about human beings to anchor their judgments [3,14]. During the laddering interview, this led them to anthropomorphize the robots based on their relatively simple behaviors (cf. [18,47]). The participants seemed unaware of their own judgment errors because of their lack of knowledge about robots [12,26,48,49]. It is not surprising then that the participants' cognitive system, which was adapted to a human environment, would produce and fail

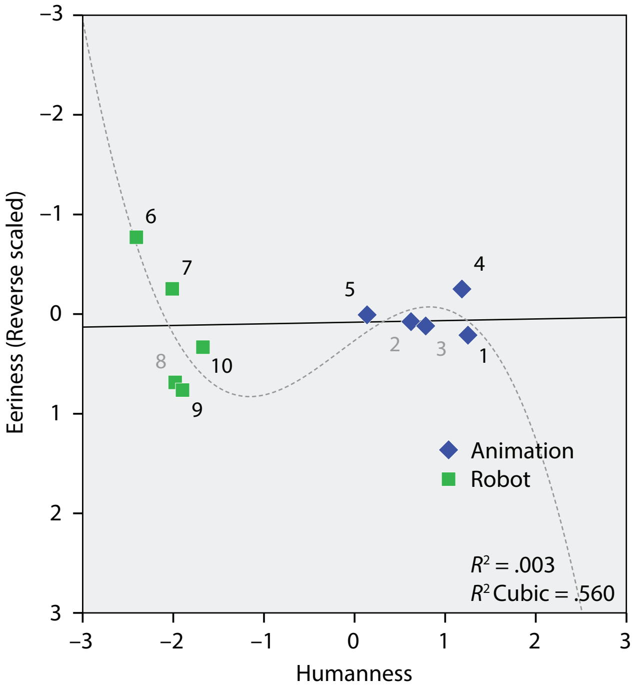

Fig. 4 Although Ho and MacDorman's [22] humanness and eeriness indices had excellent (Cronbach's $\alpha = .92$) and acceptable (.74) internal reliability, respectively, and nonsignificant linear correlation ($r = .02$, $p = .514$, straight line), the Animation and Robot groups were tightly clustered and widely separated from each other, and the Human group was omitted. (Characters 1, 4–7, 9, and 10 were used in both [22] and this study.)

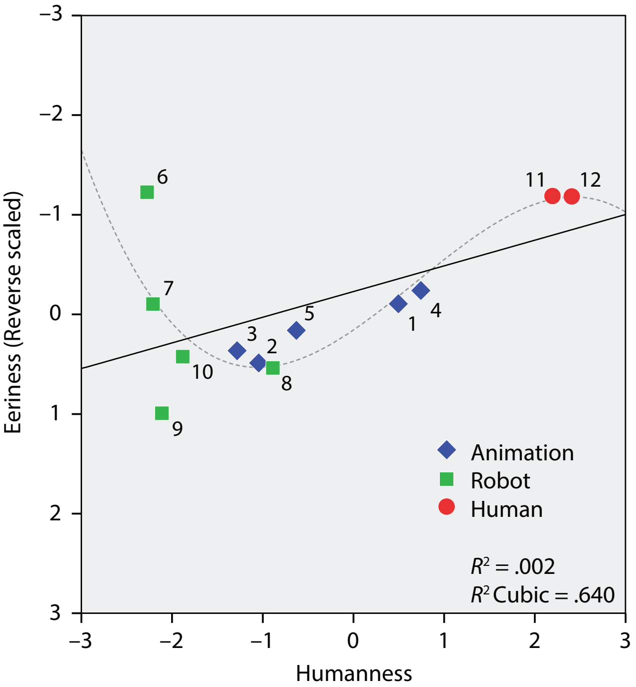

Fig. 5 The revised humanness and eeriness indices had good internal reliability (Cronbach's $\alpha = .87$ and .86, respectively) and nonsignificant linear correlation ($r = .04$, $p = .285$, straight line). The Animation and Robot were spread out and overlapped. A cubic approximation of the relation between humanness and reverse-scaled eeriness resembles Mori's (1970/2012) graph of the uncanny valley ($R^2 = .640$, dashed line).

to detect judgment errors when they were observing nonhuman, humanlike agents [1,57].

The new scales for the revised humanness, attractiveness, and eeriness indices were derived in part from the participants' responses. These adjectives may better reflect contemporary U.S. English usage and provide better content validity than previously used adjectives. The revised indices exhibited high internal reliability and, for both the computer-animated characters and robots, the bipolarity of the semantic space [4,15,28,52,59].

Confirmatory factor analysis verified the theoretical structure of the three indices, which were found to measure their putative concepts. The two subscales of the eeriness index provided a more detailed characterization of the eeriness concept. Relative to the animated characters, the robots rated higher on the eerie subscale but lower on the spine-tingling subscale.

4.1 Limitations

Although eeriness was not significantly correlated with humanness or attractiveness, attractiveness was significantly correlated with humanness with a medium effect size ($r = .36$, $p < .001$). This constitutes a substantial reduction in effect size ($r = .61$, $p < .001$) from Ho and MacDorman, Table 7 [22]. One source of correlation may be the lack of stylish mechanical-looking robots and cartoon characters among the stimuli. Nevertheless, the difficulty in decorrelating measures of attractiveness and other positive attribute dimensions from those of humanness indicates a general preference in U.S. culture for human attributes relative to nonhuman attributes and also for attractive attributes relative to unattractive attributes.

From the perspective of index development, emotional responses to robots and animation vary considerably between observers. These individual differences complicate the development of quantitative measures of the uncanny valley; thus, their effects require further investigation [7,31].

Although neither age nor gender were significant factors in our undergraduate population, these variables may become significant in a sample with a more widely distributed age range. Cultural differences and exposure can significantly affect attitudes toward robots [35]. Thus, the revised indices should be tested with with other populations (e.g., [11]).

5 Conclusion

The revised indices developed in this study have two additional advantages over their previous versions (compare Figure 3 and 5 of this study with Figure 8 and 9 of [22], respectively; Figure 9 was reproduced as Figure 4 in this paper for ease of comparison). First, the scales of each index exhibit a broader conceptual coverage; they are well differentiated from each other while, nevertheless, remaining reliable (Figure 3). Second, the humanlike characters no longer form two tightly clustered, but widely separated, categories; instead, they show considerable spread and differentiation along the humanness and eeriness dimensions — and in a *U*-shaped pattern that somewhat resembles Mori's original uncanny valley graph (Figure 5).

The revised indices also retained three advantages of the original indices. First, they maintained their theoretical structure and psychometric properties in large-scale testing. Second, their internal reliability remained high. Third, two subscales of the revised eeriness index, namely, eerie and spine-tingling, continued to serve as two stand-alone concepts for the measurement, as was verified by confirmatory factor analysis. Owing to the above advantages, these indices can contribute to the measurement and plotting of human perceptions of humanlike characters, thus providing valuable feedback to enhance their designs.

The revised indices reliably measure fairly independent dimensions with respect to the perceptions of anthropomorphic characters. In addition to assisting robot developers [2], the revised indices can also assist animators. Comparing different characters or comparing different feature settings and configurations for the same character using the same set of indices will help engineers and animators make design decisions.

Acknowledgements The authors would like to thank Debaleena Chattopadhyay, Alexander Fedorikhin, Edgar Huang, Wade J. Mitchell, Himalaya Patel, and Mark Pfaff for their insightful comments on an earlier draft of the manuscript and Ryan Sukale for technical help. This research was supported by the US National Institutes of Health (P20 GM066402) and an IUPUI Signature Center.

References

1. Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110*(3), 486–498. doi:10.1037/0033-2909.110.3.486
2. 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. doi:10.1007/s12369-008-0001-3
3. Becker-Asano, C., Ogawa, K., Nishio, S., & Ishiguro, H. (2010). Exploring the uncanny valley with Geminoid HI-1 in a real-world application. In K. Blashki (Ed.), *Proceedings of IADIS International Conference Interfaces and Human Computer Interaction* (pp. 121–128). Lisbon, Portugal: IADIS Press.
4. Bentler, P. M. (1969). Semantic space is (approximately) bipolar. *Journal of Psychology, 71*(1), 33–40. doi:10.1080/00223980.1969.10543067
5. Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 170*(2), 238–246. doi:10.1037/0033-2909.107.2.238
6. Burleigh, T. J., & Schoenherr, J. R. (2015). A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization? *Frontiers in Psychology, 5*(1488), 1–19. doi:10.3389/fpsyg.2014.01488
7. Chattopadhyay, D., & MacDorman, K. F. (2016). Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of Vision, 18*.
8. Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., & Jäncke, L. (2013). Category processing and the human likeness dimension of the uncanny valley hypothesis: Eye-tracking data. *Frontiers in Psychology, 4*(108), 1–12. doi:10.3389/fpsyg.2013.00108
9. Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": Behavioral and functional MRI findings. *Frontiers in Human Neuroscience, 5*(125), 1–14. doi:10.3389/fnhum.2011.00126
10. 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. doi:10.2307/249690
11. Destephe, M., Brandao, M., Kishi, T., Zecca, M., Hashimoto, K., & Takanishi, A. (2015). Walking in the uncanny valley: The importance of attractiveness on the acceptance of a robot as a working partner. *Frontiers in Psychology, 6*, 204. doi:10.3389/fpsyg.2015.00204
12. 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. doi:10.1111/1467-8721.01235
13. Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*(4), 752–782. doi:10.1.1.211.3309
14. 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. doi:10.1287/mnsc.1050.0409
15. 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. doi:10.1111/j.1467-9450.1976.tb00248.x
16. 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.
17. 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. doi:10.1080/10705519609540030
18. Goetz, J., Kiesler, S., and Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication* (pp. 55–60). Piscataway, NJ: IEEE Press. doi:10.1109/ROMAN.2003.1251796
19. Hanson, D. (2005). Expanding the aesthetic possibilities for humanoid robots. *Proceedings of the Views of the Uncanny Valley Workshop.* IEEE-RAS International Conference on Humanoid Robots. December 5, Tsukuba, Japan.
20. 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.
21. Hashimoto, T., Nakane, H., & Kobayashi, H. (2013). Android patient robot simulating depressed patients for diagnosis training of

psychiatric trainees. In *Proceedings of the Second IEEE International Conference on Robot, Vision and Signal Processing* (pp. 247–252). Piscataway, NJ: IEEE Press. doi:10.1109/RVSP.2013.63

22. 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. doi:10.1016/j.chb.2010.05.015
23. 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. In *Proceedings of the Third ACM/IEEE International Conference on Human–Robot Interaction* (pp. 169–176). New York, NY: ACM. doi:10.1145/1349822.1349845
24. Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology, 6*(390), 1–16. doi:10.3389/fpsyg.2015.00390
25. 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. doi:10.1177/0956797610371964
26. 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. doi:10.1037/0022-3514.77.6.1121
27. 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. doi:10.1177/0956797610388044
28. Lorr, M. & Wunderlich, R. A. (1988). A semantic differential mood scale. *Journal of Clinical Psychology, 44*(1), 33–36.
29. Ludemann, P. & Nelson, C. A. (1988). The categorical representation of facial expressions by 7-month-old infants. *Developmental Psychology, 24*(4), 492–501. doi:10.1037/0012-1649.24.4.492
30. MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition, 146*, 190–205 doi:10.1016/j.cognition.2015.09.019
31. MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies, 16*(2), 141–172. doi:10.1075/is.16.2.01mac
32. 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. doi:10.1016/j.chb.2008.12.026.
33. MacDorman, K. F., & Ishiguro, H. (2006a). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies, 7*(3), 297–337. doi:10.1075/is.7.3.03mac
34. MacDorman, K. F., & Ishiguro, H. (2006b). Opening Pandora's uncanny box: Reply to commentaries on "The uncanny advantage of using androids in social and cognitive science research." *Interaction Studies, 7*(3), 361–368. doi:10.1075/is.7.3.10mac
35. 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. doi:10.1007/s00146-008-0181-2
36. Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*(1), 93–120. doi:10.1146/annurev.psych.51.1.93
37. Mangan, B. B. (2015). The uncanny valley as fringe experience. *Interaction Studies, 16*(2), 193–199. doi:10.1075/is.16.2.05man
38. Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition, 146*, 22–32. doi:10.1016/j.cognition.2015.09.008
39. Meah, L. F. S., & Moore, R. K. (2014). The uncanny valley: A focus on misaligned cues. In M. Beetz, B. Johnston, & M.-A. Williams (Eds.), *Social Robotics* (LNAI, vol. 8755, pp. 256–265). Cham, Switzerland: Springer. doi:10.1007/978-3-319-11973-1_26
40. Michalowski, M. P., Sabanovic, S., & Simmons, R. (2006). A spatial model of engagement for a social robot. In *Proceedings of the Ninth IEEE International Workshop on Advanced Motion Control* (pp. 762–767). Piscataway, NJ: IEEE Press. doi:10.1109/AMC.2006.1631755
41. Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. *Minds and Machines, 19*(3), 345–459. doi:10.1007/s11023-009-9158-2
42. 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. doi:10.1068/i0415
43. Moore, R. K. (2012). A Bayesian explanation of the 'uncanny valley' effect and related psychological phenomena. *Scientific Reports, 2*(864), 1–5. doi:10.1038/srep00864
44. Mori, M. (2012). The uncanny valley (K. F. MacDorman & N. Kageki, Trans.). *IEEE Robotics and Automation, 19*(2), 98–100. (Original work published in 1970). doi:10.1109/MRA.2012.2192811
45. 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. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication* (pp. 35–40). Piscataway, NJ: IEEE Press.
46. Nomura, T. & Kanda, T. (2016). Rapport–expectation with a robot scale. *International Journal of Social Robotics, (8)*1, 21–30. doi:10.1007/s12369-015-0293-z
47. Prakash, A., & Rogers, W. A. (2015). Why some humanoid faces are perceived more positively than others: Effects of human–likeness and task. *International Journal of Social Robotics, 7*(2), 309-0331. doi:10.1007/s12369-014-0269-4
48. Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences, 11*(1), 37–43. doi:10.1016/j.tics.2006.11.001
49. 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. doi:10.1177/0146167202286008
50. 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* (pp. 8–13). IEEE-RAS International Conference on Humanoid Robots. December 5, Tsukuba, Japan.
51. Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009). Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. *Proceedings of the Third International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–6). Amsterdam, September 10–12. doi:10.1109/ACII.2009.5349423
52. Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*7(4), 283–294. doi:10.1037/h0026086
53. 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. doi:10.1111/1468-0394.00045
54. 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. doi:10.1162/pres.16.4.337
55. ter 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. doi:10.1016/S0167-8116(97)00029-3
56. Tondu, B., & Bardou, N. (2011). A new interpretation of Mori's uncanny valley for future humanoid robots. *International Journal of Robotics and Automation, 26*(3), 337–348. doi:10.2316/Journal.206.2011.3.206-3348

57. Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies, 8*(3), 501–517. doi:10.1075/is.8.3.11tur
58. 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. doi:10.1007/s11846-009-0034-5
59. 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. doi:10.1177/014662169401800201
60. Vanden Abeele, P. (1992). *A means-end study of dairy consumption motivation.* Report for the European Commission, EC Regulation 1000/90–43 ST.
61. Vlachos, E., & Scharfe, H. (2015). An open-ended approach to evaluating android faces. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 746–751). Piscataway, NJ: IEEE Press.
62. Walters, M., Marcos, S., Syrdal, D. S., & Dautenhahn, K. (2013). An interactive game with a robot: People's perceptions of robot faces and a gesture-based user interface. In *Proceedings of the Sixth International Conference on Advanced Computer-Human Interactions* (pp. 123–128). Lisbon, Portugal: IARIA Press.
63. Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. *Japanese Psychological Research, 55*(1), 20–32. doi:10.1111/j.1468-5884.2012.00538.x
64. 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. doi:10.1037/0278-7393.31.3.538

Chin-Chang Ho is a user experience researcher. He received his Ph.D. in Human–Computer Interaction from Indiana University in 2015. His research interests include human–computer interaction, the uncanny valley, and virtual, augmented, and mixed reality.

Karl F. MacDorman is Associate Dean of Academic Affairs at the School of Informatics and Computing, Indiana University. He was previously an Associate Professor at the School of Engineering, Osaka University. He received his Ph.D. in Computer Science from the University of Cambridge in 1997 and his B.A. in Computer Science from University of California, Berkeley in 1988. He has published more than 100 papers in human–robot interaction, machine learning, and cognitive science. His research interests include symbol grounding and symbol emergence, cognitive developmental robotics, and android science.