The Robotic Social Attributes Scale (RoSAS): Development and Validation

Colleen M. Carpinella
Disney Research
1401 Flower Street
Glendale, CA 91201
+1 (818) 553-6103
colleen.carpinella@disneyresearch.com

Alisa B. Wyman
Disney Research
1401 Flower Street
Glendale, CA 91201
+1(760) 716-2965
alisa.wyman@disneyresearch.com

Michael A. Perez
Disney Research
1401 Flower Street
Glendale, CA 91201
+1 (626) 378-6747
michael.perez@disneyresearch.com

Steven J. Stroessner
Disney Research
1401 Flower Street
Glendale, CA 91201
+1 (818) 553-4388
steve.stroessner@disneyresearch.com

ABSTRACT
Accurately measuring perceptions of robots has become increasingly important as technological progress permits more frequent and extensive interaction between people and robots. Across four studies, we develop and validate a scale to measure social perception of robots. Drawing from the Godspeed Scale and from the psychological literature on social perception, we develop an 18-item scale (The Robotic Social Attribute Scale; RoSAS) to measure people's judgments of the social attributes of robots. Factor analyses reveal three underlying scale dimensions—warmth, competence, and discomfort. We then validate the RoSAS and show that the discomfort dimension does not reflect a concern with unfamiliarity. Using images of robots that systematically vary in their machineness and gender-typicality, we show that the application of these social attributes to robots varies based on their appearance.

CCS Concepts
• **Human-centered computing~HCI design and evaluation methods** • **Applied computing~Psychology** • Computing methodologies~Philosophical/theoretical foundations of artificial intelligence

Keywords
Anthropomorphism; Human Factors; Human-Robot Interaction; Measurement; Perception; Personality Perception; Psychometric Scale; Robots; Robotics; Social Perception; Social Robots; Social Robotics

HRI '17, March 06 - 09, 2017, Vienna, Austria

ACM 978-1-4503-4336-7/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2909824.3020208

1. INTRODUCTION
People interact with machines with ever-increasing frequency, and these exchanges also play increasingly important roles in peoples' lives. Recent research has centered on the social psychology of human-machine interactions, focusing primarily on issues of anthropomorphization and the attribution of the theory of mind to machines. Less work has systematically examined how traits and characteristics associated with robots vary based on features or perceived social category membership. To do so, a psychometrically valid, standardized measure of the social attributes that people ascribe to machines must be developed. The current research offers one such measure, the Robotic Social Attributes Scale (RoSAS). The aim of this research is to offer a means to assess the central attributes implicated in human perception of robots and, ultimately, to provide the robotic community with a tool to determine how perceived attributes affect the quality of interaction with robots.

1.1 Social Perception of People
Literature in social psychology has established two universal dimensions of person perception—warmth and competence [1]. These two central dimensions are thought to reflect questions concerning basic survival—whether another person intends to help or to harm us and if they have the ability to do so.

In general, when people are evaluated as warm and competent they are seen more favorably and experience more positive interactions. Warmth judgments are commonly rendered before competence judgments and carry more weight in interpersonal interactions. For example, when presented with faces of unknown individuals, perceivers render trustworthiness judgments before competence judgments [2]. Additionally, different combinations of warmth and competence judgments elicit distinct emotions. For instance, persons perceived as high on warmth but low on competence elicit pity or sympathy, whereas the opposite combination, (low on warmth but high on competence), elicits envy or jealousy [3]. Nevertheless, warmth and competence are the main drivers of the impression formation process in judgments of humans.

Warmth has been shown to be a focal underlying dimension of robot perception as well [4]. However, additional research is warranted to examine what attributes underlie the general perception of robot warmth and how they relate to other underlying perceptual dimensions such as competence.

1.2 Social Categorization of Robots

Existing research indicates social categorization processes underlying person perception can generalize to robots. Without being encouraged to do so, people spontaneously make social category judgments of humanoid robots' gender, race, and nationality [5]. People appear to apply gender categories and respective stereotypes to humanoid robots depending on whether their hairstyle implies a female or male gender identity [6]. In a similar study, female robots with long hair and full lips were perceived as higher on the communal dimension (e.g., friendly, polite, affectionate), whereas the male robots were perceived as more agentic (e.g., assertive, determined, authoritative) [7]. Therefore, social attributions typically applied to humans can also be applied to robots, especially if they are classified into gender categories. Moreover, the tendency to use gender stereotypes extends to other kinds of machines [8, 9].

People often anthropomorphize robots, imbuing them with human traits, goals, and motivations. The tendency to assign human characteristics to robots varies based on social contextual cues as well as people's motivations. People are more likely to anthropomorphize robots when human characteristics are accessible and applicable (e.g., when they have recently been used or when robots appear more humanlike) [10]. People are more likely to anthropomorphize a robot when the robot is perceived as an in-group member, sharing a common identity with them, compared to when the robot is perceived as an out-group member [11]. This suggests that the perceived anthropomorphic qualities of a robot are malleable based on characteristics central to the perceiver and the robot.

1.3 Measurement of Robot Characteristics

There is a wide-spread need for a standardized measurement tool in HRI research to allow comparisons across robots and across studies. The Godspeed scale [12] was developed as a way to measure human and robotic interaction, and it has become a widely used measure of human-robot interaction [13]. There are five central dimensions to the Godspeed scale: i) anthropomorphism, or the extent to which a robot appears humanlike versus machinelike [14]; ii) animacy, or how lifelike a robot seems [15, 16]; iii) likeability, or how friendly a robot seems [14]; iv) perceived intelligence of the robot [17, 18]; and v) perceived safety, or emotional state/anxiety of the perceiver [19].

Despite the widespread use of the Godspeed scale, little empirical work has examined its psychometric properties. Indeed, recent research has documented some shortcomings of the Godspeed indices [4]. Our work, presented here, and that of Ho and MacDorman [4] identifies several problematic aspects of the Godspeed scale. First, several of the scale items are confounded with positive and negative affect. Second, the items do not load as expected onto the five scale dimensions and several items do not load onto any dimensions. In other words, the items do not correspond to the underlying constructs they are meant to measure. Third, four out of the five dimensions are highly correlated, suggesting that similar rather than distinct concepts are measured. Finally, the semantic differential response format is used (i.e., the two endpoints of the scale are polar adjectives). Whereas some scale items use antonyms as endpoints, allowing a clear identification of the underlying construct being measured (unfriendly-friendly), other pairings appear to reflect more than one dimension of judgment (awful-nice). Our research uses items from the Godspeed and also insights from the psychological literature on social perception to develop and validate a more psychologically valid scale of robotic perception.

2. STUDY 1

Study 1 offers an exploratory factor analysis of responses to the Godspeed Questionnaire to assess the psychometric properties of the scale. As described earlier, the Godspeed attempts to measure five distinct dimensions: *anthropomorphism* (fake–natural, machinelike–humanlike, unconscious–conscious, artificial–lifelike, moving rigidly–moving elegantly), *animacy* (dead–alive, stagnant–lively, mechanical–organic, artificial–lifelike, inert–interactive, apathetic–responsive), *likeability* (dislike–like, unfriendly–friendly, unkind–kind, unpleasant–pleasant, awful–nice), *perceived intelligence* (incompetent–competent, ignorant–knowledgeable, irresponsible–responsible, unintelligent–intelligent, foolish–sensible), and *perceived safety* (anxious–relaxed, agitated–calm, quiescent–surprised). Using these semantic differential scale items included in the original Godspeed Scale, we tested whether the hypothesized 5-factor solution would emerge and whether items would load on the expected dimensions. Our study used 23 of the 24 Godspeed items, because artificial–lifelike appears on both the *anthropomorphism* and *animacy* subscales of the Godspeed.

2.1 Study 1 Method

2.1.1 Participants

Two hundred fifteen people (127 men, 88 women) participated via Amazon's Mechanical Turk (mTurk) in exchange for $0.60.

2.1.2 Procedure

We informed participants that we were interested in how people perceive groups in our society. Participants were told that when we encounter a name of a group, certain words might come to mind, and we were interested in measuring these associations. Participants were asked to evaluate the category 'robots' on the Godspeed items. We intentionally left the category of interest broad and did not include any images, a definition, or specification of the type of robot. We did not want the development of our scale to be tied to a specific exemplar or type of robot, so that it could serve as a general measure of robot perception.

Participants were presented with the list of attributes from the Godspeed scale in the original semantic differential format and were asked to rate the extent to which they perceived these attributes to be associated with robots. Participants were asked, "Using the scale provided, what is your impression of the category robots?" Participants responded using a 5-point likert scale for each sematic differential item. The order in which the items were presented was randomized. After providing judgments, participants were asked to fill out demographic information about themselves and thanked for their participation.

2.2 Study 1 Results

We performed an exploratory factor analysis (EFA) to assess the properties of the Godspeed scale. EFA is a statistical method used to identify conceptual variables being measured by a scale and the

relationships between variables. When using EFA, you need to specify an extraction method which determines the statistical procedures that are utilized. To identify the five factors underlying the Godspeed scale items, we used principal axis factoring and designated a five factor solution [20]. Next, we needed to pick a rotation method. We chose promax rotation because there was reason to believe that the factors would be associated with one another [20].

Next, we conducted the EFA, plotted the scree plot, and examined the eigenvalues to determine the relative importance of the factors. Eigenvalues represent how much variation in scale responding can be explained by each factor; the larger the eigenvalue, the more the factor explains. [20]. The scree plot depicted in Figure 1 shows the eigenvalues (y-axis) plotted against the factor with which they are related (x-axis). We determined the importance of the factors by examining the eigenvalues. The scree plot shows a clear leveling-off point indicating that only the first three factors are important. Altogether, these three factors accounted for 46% of the overall variance in robot evaluation. Inspection of the items that loaded on each factor (factor loadings > .500) indicated that they reflected perceived anthropomorphism, perceived intelligence, and likeability. The factor loadings are depicted in Table 1, with factor loadings > .500 bolded.

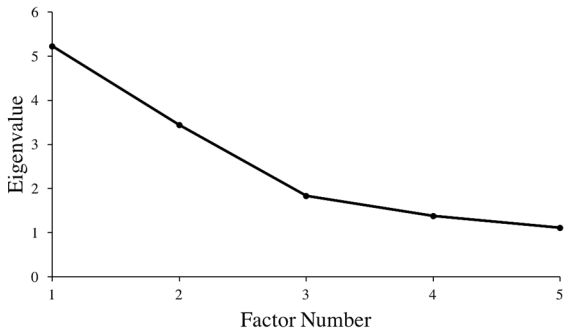

Figure 1. Scree Plot (Study 1)

The first factor reflected *anthropomorphism.* Four items comprised this factor: machinelike–humanlike, mechanical–organic, artificial–lifelike, moving rigidly–moving elegantly (α = .77). In comparing this factor against the Godspeed subscales, two items that should have loaded on this factor – fake-natural and unconscious-conscious – did not (.290 and .266, respectively), and one item that was not expected on this factor – mechanical–organic – did. While a test of the reliability for the original five items was reasonable (α = .74), it is clear that the factor that emerged here was similar to but distinct from what is reflected in the Godspeed.

The second factor reflected *perceived intelligence*. Six items comprised this factor: incompetent–competent, ignorant–knowledgeable, foolish–sensible, unintelligent–intelligent, inert–interactive, and irresponsible–responsible (α = .81). Whereas five of these items comprised the intelligence subscale of the Godspeed, one item – inert–interactive – did not. However, the reliability for the original five items was nearly identical (α = .82).

Table 1. Factor Loadings (Study 1)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Machinelike-Humanlike	**.846**	.071	-.094	-.108	.177
Mechanical-Organic	**.712**	-.133	-.140	.137	.137
Artificial-Lifelike	**.630**	.039	.058	.067	.009
Moving Rigidly-Moving Elegantly	**.527**	-.011	-.018	.101	-.041
Incompetent-Competent	-.035	**.796**	-.134	-.078	.088
Ignorant-Knowledgeable	-.167	**.747**	-.006	.215	.016
Foolish-Sensible	-.039	**.688**	.065	-.077	.092
Unintelligent-Intelligent	.014	**.564**	-.055	.214	.093
Inert-Interactive	.092	**.526**	.103	-.097	-.234
Irresponsible-Responsible	.075	**.517**	.062	.034	.096
Awful-Nice	-.004	-.050	**.792**	-.051	.118
Dislike-Like	-.081	.008	**.662**	-.028	.098
Unpleasant-Pleasant	.041	.094	**.606**	-.099	.051
Unkind-Kind	-.084	-.134	**.551**	.256	.088
Dead-Alive	.097	.121	-.044	**.592**	-.066
Unconscious-Conscious	.266	.166	-.038	**.571**	-.070
Anxious-Relaxed	.153	.064	.088	-.106	**.646**
Agitated-Calm	-.134	.131	.221	.085	**.543**
Unfriendly-Friendly	.361	-.010	.493	-.063	-.066
Stagnant-Lively	.356	.171	.150	.010	-.149
Quiescent-Surprised	.333	-.065	.129	.147	-.226
Fake-Natural	.290	-.207	.077	.449	.077
Apathetic-Responsive	.012	.197	.337	.062	-.191
Eigenvalue	5.224	3.440	1.830	1.378	1.109
Percent variance explained	22.71%	14.96%	7.96%	6.00%	4.82%

The scree plot suggested that Factors 4 and 5 should not be included in the scale. However, we explored these factors to determine whether they corresponded with the animacy and perceived safety components of the Godspeed. The two items that load on Factor 4 are dead–alive and unconscious–conscious. Only the first appears in the animacy subscale of the Godspeed, and the second is associated with anthropomorphism in the Godspeed. None of the other five items from the animacy subscale of the Godspeed loaded on Factor 4 (all loadings < .137). Moreover, the reliability for the original six animacy items is low (α = .57). Therefore, it appears that Factor 4 bears little resemblance to the Godspeed animacy subscale.

Factor 5 possibly reflects perceived safety as evidenced by anxious–relaxed and agitated–calm. However, the item quiescent-surprised actually loaded negatively on this factor (-.226). Moreover, the reliability for the original three perceived safety items is unacceptably low (α = .22). Therefore, based upon the scree plot and the low overlap between these results and the constructs reflected in the Godspeed, it does not appear that the proposed 5-factor solution offers a psychometrically sound characterization of scale responses.

Table 2. Factor Correlation Matrix (Study 1)

Factor	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1				
Factor 2	-.01			
Factor 3	.27	.47		
Factor 4	.41	.25	.25	
Factor 5	-.09	.20	.23	.10

In examining the correlations between the factors, the first three were relatively independent with the exception of the correlations between perceived intelligence and likeability (r = .47) and anthropomorphism and Factor 4 (r = .47) (see Table 2). These results suggest that some of the Godspeed subscales might be measuring related rather than independent constructs of robot perception.

In summary, Study 1 shows that the items in the Godspeed scale load onto three unique factors— reflecting anthropomorphism, perceived intelligence, and likeability Unexpectedly, factors reflecting animacy and perceived safety did not emerge as strong constructs utilized in judging robots as a social category.

3. STUDY 2

In Study 2, we sought to complement the items from the Godspeed scale with a novel set of social attributes generated from psychological literature on social cognition. To the degree that robots are often anthropomorphized, we hypothesized that attributes central to social judgment might also play an important role in robot perception. Therefore, we aimed to develop a psychometrically valid scale to evaluate robots on social dimensions (i.e., the Robotic Social Attribute Scale; RoSAS). Participants in Study 2 were presented with the Godspeed items and a list of attributes shown to be central in social perception: The Stereotype Content Model and the Bem Sex Role Inventory [1, 21]. We sought to determine the underlying factor structure and to examine the overlap with the 3-factor solution derived from assessing responses using the Godspeed scale.

3.1 Study 2 Method

3.1.1 Participants

Two hundred ten people (105 women, 104 men, 1 unidentified) participated via mTurk in exchange for $1.25.

3.1.2 Procedure

Participants were presented with modified items from the Godspeed scale and a set of attributes from research on social cognition. To make items more comparable, we separated out the endpoints from the semantic differential Godspeed items to create unidimensional items. This generated a list of 38 unique items from the Godspeed. We then combined these items with our attributes, eliminating duplicate items, yielding a total of 83 items. Participants were asked, *"Using the scale provided, how closely are the words below associated with the category robots?"*. Participants responded using a 9-point likert scale from 1 = *definitely not associated* to 9 = *definitely associated*. The order in which the items were presented was randomized.

3.2 Study 2 Results

We performed an exploratory factor analysis on participants' responses in an attempt to develop a broader and more psychometrically valid measure of robot perception. We used principal axis factoring extraction method with a promax rotation. This factor analysis revealed three factors with eigenvalues greater than one, and the scree plot showed a clear leveling-off point after the third factor (see Figure 2). Together, these three factors accounted for 44% of the overall variance in robot evaluation. Inspection of the items that loaded on each factor (factor loadings > .500) indicated that 18 items loaded onto the three factors (see Table 3). There was no cross loading of any of the unlisted items.

We labeled the first factor *warmth.* The items that comprised this factor were: feeling, happy, organic, compassionate, social, and emotional (α = .91). The items that loaded onto the second factor related to the intelligence or ability of the robot, and we labeled this factor *competence.* The items that comprised this factor were: knowledgeable, interactive, responsive, capable, competent, and reliable (α = .84).

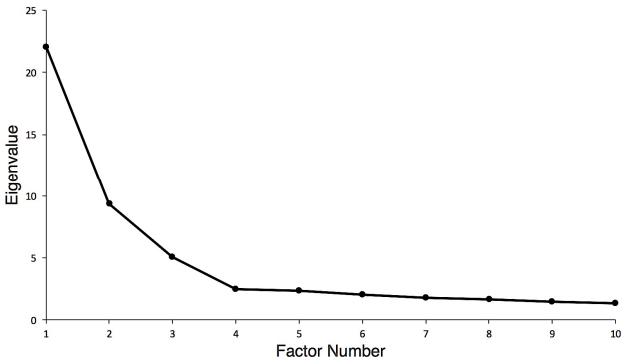

Figure 2. Scree Plot (Study 2)

Factor 3 items were related to awkwardness, and we labeled this factor *discomfort*. The items that comprised this factor were aggressive, awful, scary, awkward, dangerous, and strange (α = .82). The correlations between the factors were lower than for those between the Godspeed subscales alone (see Table 4). In examining the correlations between the factors, the factors were relatively independent (see Table 4), suggesting that the three dimensions of the RoSAS are measuring independent constructs of robot perception.

Therefore, the factor analysis for the combined Godspeed items and attributes from the social cognition literature indicates that there are three central factors underlying the evaluation of robots. The first two dimensions of warmth and competence are similar to the two trait dimensions central to person perception [1], and parallel the likeability and perceived intelligence dimensions from the Godspeed items. This attests that these are two central dimensions underlying robot evaluation. Furthermore, a third aspect of discomfort arises with the combined sets and somewhat parallels aspects of the Godspeed analysis (i.e., a thematically coherent but weak 5th factor reflecting safety concerns).

Table 3. Factor Loadings (Study 2)

Variable	Factor 1	Factor 2	Factor 3
Happy	**.831**	-.013	-.009
Feeling	**.811**	-.154	.043
Social	**.793**	.125	-.176
Organic	**.784**	-.149	-.022
Compassionate	**.778**	-.039	.030
Emotional	**.776**	-.204	.050
Capable	-.210	**.706**	-.052
Responsive	-.141	**.680**	.040
Interactive	-.225	**.652**	-.006
Reliable	-.061	**.651**	-.028
Competent	-.111	**.646**	-.040
Knowledgable	-.007	**.620**	-.021
Scary	.052	-.012	**.693**
Strange	-.053	.150	**.601**
Awkward	.049	.037	**.601**
Dangerous	-.035	.024	**.597**
Awful	.360	-.250	**.555**
Aggressive	.265	.009	**.547**
Eigenvalue	22.031	9.336	5.047
Percent variance explained	26.54%	11.25%	6.09%

The combined analysis suggests that the scale developed here has stronger psychometric properties compared to the Godspeed scale. Eigenvalues are higher (all > 5.5 vs. < 5.5, respectively) and scale reliabilities are higher (all > .82 vs. all < .82, respectively) for the combined set of attributes. Taken together, we believe that these 18 items – that we call the Robotic Social Attributes Scale (RoSAS) – offer a parsimonious and psychometrically valid scale for the social evaluation of robots.

Table 4. Factor Correlation Matrix (Study 2)

Factor	Factor 1	Factor 2
Factor 1		
Factor 2	.20	
Factor 3	.34	.18

4. STUDY 3

Study 3 focused on the unique attribute of *discomfort* that emerged in the factor analysis. This factor does not appear in measures of social perception of humans, and we considered two possible reasons why it might play a role in robot perception. One possibility is that people are concerned about feelings of discomfort that might arise in interacting with robots. A second possibility is that people might be concerned about discomfort in considering any entity that is unfamiliar. Thus, Study 3 was designed to test whether the unique third dimension of discomfort that emerged in our study of robot perception arose due to the confounding factor of unfamiliarity. In other words, perhaps the presence of a *discomfort* dimension was simply due to a lack of familiarity with robots. If so, a 3-dimension solution including a factor focusing on discomfort might emerge when people judge any unfamiliar entity. If discomfort reflects a concern with unfamiliarity, then a 3-factor solution should fit the data for novel but not for familiar entities.

4.1 Study 3 Method

4.1.1 Participants

Seventy people (36 men, 33 women, 1 unidentified) participated via mTurk in exchange for $1.20.

4.1.2 Procedure

Participants were presented with familiar and unfamiliar animal (i.e., giraffe and okapi) and human (i.e., Australian and Nauruan) linguistic categories. Participants then rated each entity on the 18-items of RoSAS. Participants were asked, "Using the scale provided, how closely are the words below associated with the category [giraffe, okapi, Australian, and Nauruan]?" Participants responded using a 9-point likert scale from 1= *definitely not associated* to 9 = *definitely associated*. The order in which the items were presented was randomized.

4.2 Study 3 Results

We performed an exploratory factor analysis on participants' responses separately for each animal and human category that was provided. We used principal axis factoring extraction method with a promax rotation.

The items for all four categories loaded onto 2 factors. For all four categories, the eigenvalues for the first two factors were above 3, while the eigenvalues for the remainder of the factors were below 2. This pattern of results held for the familiar and unfamiliar animal categories. For giraffe, the first and second factors (6.19 and 3.14) had higher eigenvalues than the third and fourth factors (1.62 and 1.26). The pattern was the same for okapi with a clear divide between the eigenvalues for the first and second (7.25 and 3.02) and the third and fourth factors (1.32 and 1.19). The familiar and unfamiliar human categories showed the same pattern. Aussie showed a clear divide between the eigenvalues for the first and second (7.38 and 3.21) and the third and fourth factors (1.13 and 1.04). For Nauruan, the first and second factors (7.97 and 3.90) had higher eigenvalues than the third and fourth factors (1.23 and 0.88). Finally, the scree plots for all four categories showed a clear leveling-off point after the second factor, indicating that for both familiar and unfamiliar entities, no discomfort dimension arises.

Therefore, the emergence of a factor focusing on discomfort in judgments of robots does not appear to arise simply due to a lack of familiarity. Studies 2 and 3 provided evidence that people spontaneously consider warmth and competence in thinking about robots, just as they do when they think about people. However, people appear to spontaneously consider discomfort when judging robots but do not do so when thinking of other unfamiliar groups. In Study 4, we sought to further validate RoSAS by testing whether the judgment of robots on these social attributes varies as a function of robot appearance.

5. STUDY 4

One implication of RoSAS is that people think about robots using dimensions of assessment that are central to judgments of human beings, specifically focusing on issues of warmth and competence. We wondered whether these attributes, which are associated with the important human social category of gender (i.e., gender stereotypes), would influence judgments of robots that had male and female features.

A second implication of RoSAS is that people focus uniquely on discomfort when thinking about robots compared with humans. We wondered whether inferences about discomfort would differentially appear for male versus female robots. We also wanted to simultaneously test how robots that varied along the humanlike-machinelike dimension would be perceived. Therefore, Study 4 tested whether robot faces that varied systematically in the gender-typicality and machine-human quality of their appearance, would be perceived differently in terms of warmth, competence, and discomfort.

5.1 Study 4 Method

5.1.1 Participants

Two hundred fifty-two people (137 women, 115 men) participated via mTurk in exchange for $1.80.

5.1.2 Stimuli and Procedure

The base faces that were used to create the robot stimuli were generated using commercial software (FaceGen Modeler) [22]. Images were Caucasian individuals manipulated to appear feminine, androgynous, or masculine. After these base faces were created, the images were edited in Photoshop to appear humanlike, human-machine blend ("blended"), or machinelike. The facial structure was kept constant, but the texture of the skin was systematically and consistently varied depending on the target type, with machinelike images constructed with a metallic skin and visible hardware. The full set of robot faces varied systematically in terms of their gender-typicality (3 feminine, 3 androgynous, 3 masculine) and their machineness (3 humanlike, 3 human-machine blend, 3 machinelike) (see Figure 3). Participants were informed that all images they were robots and in debriefing, no participants expressed concern that they were not. The gender-typicality and machineness of the images was not labeled.

Stimuli were pretested to ensure that the machineness and gender-typicality of the images was manipulated accordingly. Pretest data indicated that the machinelike stimuli were indeed rated significantly higher on machineness (M = 6.2, SD = 1.1) than blended (M = 4.9, SD = 1.3) and humanlike stimuli (M = 2.6, SD = 1.3), $F(2,122) = 238.20$, $p < .001$. Pretest data also showed that male stimuli were rated significantly higher on masculinity (M = 6.1, SD = .9) compared to the androgynous (M = 5.0, SD = 1.0) and female stimuli (M = 2.8, SD = .9), $F(2,122) = 45.20$, $p < .001$.

Participants were randomly presented with each of the nine robot images (see Figure 3). Participants completed the RoSAS for each robot face following the procedures discussed earlier. The presentation of the 18 traits was randomized across participants.

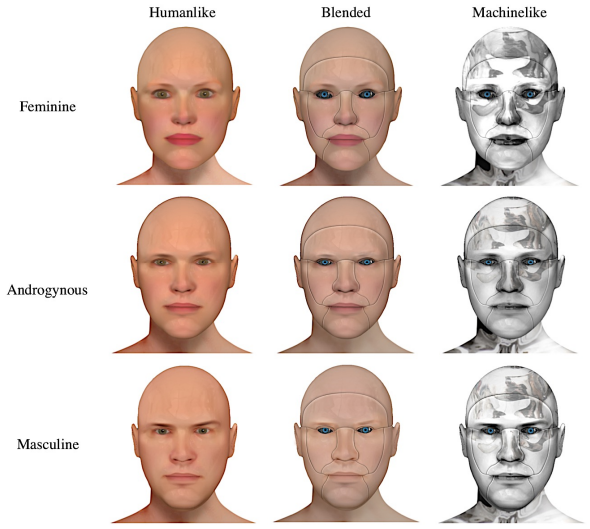

Figure 3. Robot Image Stimuli (Study 4)

5.2 Study 4 Results

5.2.1 Analytic Strategy

Image judgments were nested under participant. We therefore analyzed data using generalized estimating equations to accurately model the hierarchical nature of the data [23], specifying a normal distribution. We report unstandardized regression coefficients (B). Target Gender (1 = Female, 2 = Androgynous, 3 = Male), and Target Type were coded multicategorically (1 = Human, 2 = Human-Machine, 3 = Machine). All analyses were run in a stepwise fashion, first testing main effects and subsequently adding predicted interactions to the model. The six items comprising the warmth, competence, and discomfort dimensions were averaged. All three dimensions showed excellent reliability, $\alpha = .92$, $\alpha = .95$, and $\alpha = .90$, respectively, across all target images.

5.2.2 Warmth Judgments

Based on well-established human stereotypes, we predicted that female robots would be rated higher on warmth than male robots. To test this prediction, we regressed Warmth onto Target Gender, Target Type, and their interaction. There was a main effect of Target Gender, $\chi^2(2) = 27.80$, $p < .0001$; female robots and androgynous robots were rated as warmer than male robots, *Contrasts* = -.26 and -.20, *SE*s = .04, 95% CIs = [-.33, -.19] and [-.27, -.13]. Female robots and androgynous robots did not vary in their perceived warmth, *Contrast* = -.06, *SE* = .04, 95% CI = [-.13, .01] (see Figure 4).

We also expected that humanlike robots would be rated highest on warmth. There was a main effect of Target Type, $\chi^2(2) = 309.52$, $p < .0001$; humanlike robots were rated as significantly warmer than blended robots and machinelike robots, *Contrasts* = -.72 and -.1.03, *SE*s = .04, 95% CIs = [-.80, -.65] and [-1.10, -.96]. Blended robots were seen as significantly warmer than machinelike robots, *Contrasts* = -.31, *SE* = .04, 95% CI = [-.38, -.23]. The interaction between Target Gender and Target Type was not significant, $\chi^2(4) = 5.25$, $p = .2629$. The combination of these two main effects resulted in female humanlike robots being rated the highest on warmth.

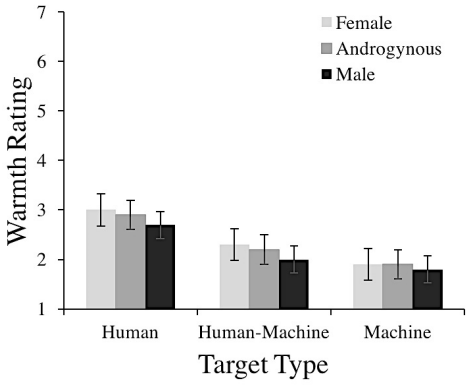

Figure 4. Warmth Judgments (Study 4)

5.2.3 Competence Judgments

Based on well-established human stereotypes, we first predicted that male robots would be rated highest on competence. To test this prediction, we regressed Competence onto Target Gender, Target Type, and their interaction. There was a main effect of Target Gender, $\chi^2(2) = 8.24$, $p = .0163$; contrary to our predictions, female robots and androgynous robots were rated as more competent than male robots, *Contrasts* = -.08 and -.14, *SE*s = .03, 95% CIs = [-.14, -.01] and [-.20, -.07]. However, female robots and androgynous robots did not vary in their perceived competence, *Contrast* = .06, *SE* = .03, 95% CI = [-.01, .12] (see Figure 5).

Next, there was a main effect of Target Type, $\chi^2(2) = 52.10$, $p < .0001$; humanlike robots were rated as significantly more competent than blended robots and machinelike robots, *Contrasts* = -.35 and -.37, *SE*s = .03, 95% CIs = [-.42, -.29] and [-44, -.31]. The perceived competence of blended and machinelike robots did not vary, *Contrast* = -.02, *SE* = .03, 95% CI = [-.09, .04].

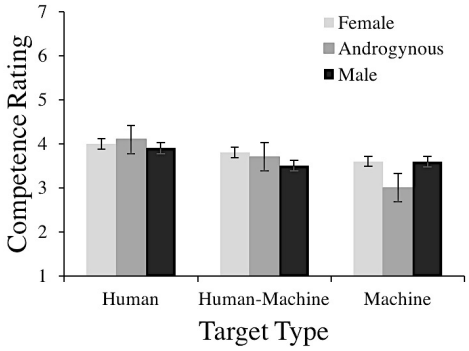

Figure 5. Competence Judgments (Study 4)

The interaction between Target Gender and Target Type was significant, $\chi^2(4) = 9.96$, $p = .0411$. In line with our predictions, robots that appeared humanlike were seen as more competent than machinelike robots. However, contrary to our predictions, female and androgynous robots were also seen as more competent.

5.2.4 Discomfort Judgments

We predicted that male robots would produce higher judgments of discomfort than female robots [5]. To test this prediction, we regressed Competence onto Target Gender, Target Type, and their interaction. There was a main effect of Target Gender, $\chi^2(2) = 24.04$, $p < .0001$; in line with our predictions, male robots and androgynous robots were rated higher on discomfort compared to female robots, *Contrasts* = .28 and .37, *SE*s = .05, 95% CIs = [.18, .37] and [.28, .47]. Male robots and androgynous robots did not vary in their perceived discomfort, *Contrast* = -.10, *SE* = .05, 95% CI = [-.19, .001] (see Figure 6).

Next, we predicted that machinelike robots would be anthropomorphized the least and therefore rated highest on discomfort. There was a main effect of Target Type, $\chi^2(2) = 283.45$, $p < .0001$; machinelike robots were rated as significantly higher on discomfort than blended robots and human robots, *Contrasts* = .49 and 1.36, *SE*s = .05, 95% CIs = [.39, .59] and [1.26, 1.46]. Blended robots were rated higher on discomfort than humanlike robots, *Contrast* = .87, *SE* = .05, 95% CI = [.77, .97]. The interaction between Target Gender and Target Type was significant, $\chi^2(4) = 11.15$, $p = .0250$. Male machinelike robots were rated the highest on discomfort.

The findings from Study 4 indicate that robots that appear female and humanlike are perceived to be warmer and more competent. Conversely, a robot appearing more male and machinelike is rated the highest on discomfort. These results show some evidence of traditional gender stereotyping and provide information about the types of robotic features that will produce more versus less discomfort.

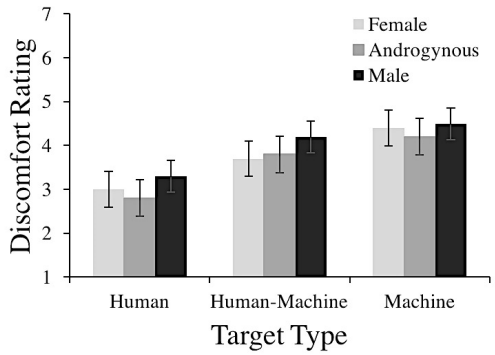

Figure 6. Discomfort Judgments (Study 4)

6. GENERAL DISCUSSION

6.1 Scale Development

Across 4 studies, we collected data with the original Godspeed items and a new set of items based on social psychological research to develop the RoSAS. We arrived at a three factor solution encompassing 18 items to characterize judgments of robots, reflecting the dimensions of warmth, competence, and discomfort. We then conceptually tested this item set with the application to familiar versus unfamiliar entities. The items loaded on only 2 factors for both familiar and unfamiliar entities, indicating that the discomfort dimension is not related to people's lack of familiarity with robots.

Although we believe the RoSAS scale is a valuable new tool for measuring HRI, we do not mean to suggest that RoSAS should completely replace existing instruments. We view the Godspeed as a complement to the RoSAS when specific constructs not well measured by the RoSAS are the focus of research. The development of the RoSAS is indebted to the Godspeed in that item variants from the Godspeed scale were used to create the RoSAS, and the Godspeed offers numerous items that that are not represented in the RoSAS. However, the current work raises questions about the psychometric properties of the Godspeed of which researchers should be aware if they use items from that instrument.

Many of the studies that have used the Godspeed scale to measure robot perception have done so with specific robots. However, the development of our scale was not tethered to specific images or videos of robots. It is precisely because the RoSAS was developed for a broad category of robots that it can serve as a standardized measurement of robot perception. We see value in developing a scale that is not normed to one particular set of robots. The type of robot that an individual considers will impact the extent to which they associate different attributes with robots; however, we aimed to identify the underlying dimensions of robot judgment. We fully expect the RoSAS ratings to differ for robots with varying features and in differing roles.

6.2 Scale Validation

Validation of RoSAS demonstrates that robots' appearance impacts perceiver's social evaluations. A robot that appears female and humanlike is perceived to be warmer and more competent; whereas, a robot that appears male and machinelike is rated highest on discomfort. Results showing that humanlike robots are evaluated more favorably in terms of their warmth and competence and lower on discomfort are not surprising given that anthropomorphic robots are viewed more positively than mechanomorphic robots [24]. Additionally, in a study examining children's perceptions of robots, they perceived humanlike robots as having feelings and being better able to understand them compared to machinelike robots [25].

These findings have implications beyond measuring responses to existing robots. We would suggest they also provide useful information about robot design and interaction. A robot that appears female and humanlike is more likely to be perceived positively (i.e., higher on warmth and competence and lower on discomfort). People may be more willing to interact with such a robot and may respond better to the robot in an interaction. Conversely, a robot intended to create fear for entertainment purposes would likely benefit from having male, machinelike features. In addition, the RoSAS can be used to determine how systematic variations in robot design might affect various aspects of experience central to human-robotic interactions.

We envision three main uses of the RoSAS. First, this scale (see Table 3 for complete set of RoSAS items) can be used as a tool to evaluate robots that have already been designed. People's perceptions of robots determine expectations that guide their robot interactions. RoSAS scores for a particular robot can inform expectations about the smoothness of social interactions. Second, the RoSAS can inform development and design of robots, especially anthropomorphic robots which mimic human appearance and behavior. For example, a goal of developing a child-friendly robot could be better achieved by testing how

design choices maximize warmth and minimize discomfort. Third, the RoSAS serves as a standardized metric in HRI. Other researchers have developed a psychological scale specially for the perception of humanoid robots [26]. However, the RoSAS provides a psychometrically validated, standardized measure that can be used to measure robots developed by different people in different places for differing purposes and over time.

6.3 Limitations and Future Directions

The manner in which people evaluate robots on social dimensions presumably influences their willingness to interact with robots. However, that was not measured in these studies. Future research should examine whether there is a direct association between higher warmth and competence ratings and lower discomfort ratings and the willingness to interact with a robot. For example, questions concerning contact intention and design preference would be a valuable contribution for the design of robots that interact with humans. While the aim of our scale development was to create a scale that is generalizable to robots that vary in their appearance and role, the application of the scale to individual robots would be an additional way to utilize the RoSAS. The consequence of this application would be to smooth human-robotic interactions, especially in contexts where robots are meant to assist individuals or to provide positive interactive experiences.

One limitation of the current research is that we validated RoSAS using images of robots which did not permit any actual robot interaction. People respond differently to an embodied robot compared to a robot image projected on a screen [27]. In order to validate the scale, it remains important to move beyond images and videos by using actual robots in real-life situations.

Future research could also further examine the role of perceiver gender in robot perception. People often feel close to and more strongly anthropomorphize same-gender robots compared to opposite gender robots [7]. Although we do not find evidence in Study 4 that perceiver gender moderates any of the reported relationships, future studies should consider participant information in determining whether RoSAS evaluations hold across different demographic groups (i.e., gender and age). Further work could also be done on the specific features that trigger gender stereotyping of robots. Given that robots can display features not possible for humans, RoSAS can be a valuable tool for understanding when feature variation elicits gender stereotypes.

7. ACKNOWLEDGMENTS

We would like to thank Corey McClelland for his help in developing the robot image stimuli used in Study 4. We would also like to thank Jonathan Benitez for his feedback on prior versions of this manuscript.

8. REFERENCES

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

[2] Willis, J., & Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592-598.

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

[4] 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.

[5] Carpenter, J., Davis, J. M., Erwin-Stewart, N., Lee, T. R., Bransford, J. D., & Vye, N. (2009). Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics*, 1(3), 261-265.

[6] Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4), 724-731.

[7] Eyssel, F., & Hegel, F. (2012). (S)he's Got the Look: Gender Stereotyping of Robots. *Journal of Applied Social Psychology*, 42(9), 2213-2230

[8] Nass, C., Moon, Y., & Green, N. (1997). Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers with Voices. *Journal of Applied Social Psychology*, *27*(10), 864-876.

[9] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81-103.

[10] Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864-886.

[11] Kuchenbrandt, D., Eyssel, F., Bobinger, S., & Neufeld, M. (2013). When a robot's group membership matters. *International Journal of Social Robotics*, *5*(3), 409-417.

[12] 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.

[13] Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., & Petrick, R. (2012). Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 3-10).

[14] Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009, September). My robotic doppelgänger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 269-276).

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

[16] Bartneck, C., Kanda, T., Mubin, O., & Al Mahmud, A. (2007, November). The perception of animacy and intelligence based on a robot's embodiment. In *2007 7th IEEE-RAS International Conference on Humanoid Robots* (pp. 300-305).

[17] Warner, R. M., & Sugarman, D. B. (1986). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology*, *50* (4), 792-799.

[18] Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1996, November). My partner is a real dog: cooperation with social

agents. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work* (pp. 399-408).

[19] Kulic, D., & Croft, E. (2005, August). Anxiety detection during human-robot interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 616-621).

[20] Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. SAGE Publications Ltd.

[21] Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Clinical and Consulting Psychology, 42,* 155-162.

[22] Blanz, V., & Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187-194). ACM Press/Addison-Wesley Publishing Co.

[23] Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. Applied longitudinal analysis. 2004. *Hoboken Wiley-Interscience*.

[24] Fraune, M. R., Sherrin, S., Sabanović, S., & Smith, E. R. (2015, March). Rabble of robots effects: Number and type of robots modulates attitudes, emotions, and stereotypes. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 109-116).

[25] Woods, S., Dautenhahn, K., & Schulz, J. (2004). The design space of robots: Investigating children's views. In Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop, 47-52.

[26] Kamide, H., Mae, Y., Kawabe, K., Shigemi, S., & Arai, T. (2012). A psychological scale for general impressions of humanoids. In *IEEE International Conference on Robotics and Automation* (pp. 4030-4037).

[27] Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition,* 26(2), 169-181.