What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database

Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle
Brown University
Providence, Rhode Island
[elizabeth_phillips1,xuan_zhao,daniel_ullman,bfmalle]@brown.edu

ABSTRACT

Anthropomorphic robots, or robots with human-like appearance features such as eyes, hands, or faces, have drawn considerable attention in recent years. To date, what makes a robot appear human-like has been driven by designers' and researchers' intuitions, because a systematic understanding of the range, variety, and relationships among constituent features of anthropomorphic robots is lacking. To fill this gap, we introduce the ABOT (Anthropomorphic roBOT) Database—a collection of 200 images of real-world robots with one or more human-like appearance features (http://www.abotdatabase.info). Harnessing this database, Study 1 uncovered four distinct appearance dimensions (i.e., bundles of features) that characterize a wide spectrum of anthropomorphic robots and Study 2 identified the dimensions and specific features that were most predictive of robots' perceived human-likeness. With data from both studies, we then created an online estimation tool to help researchers predict how human-like a new robot will be perceived given the presence of various appearance features. The present research sheds new light on what makes a robot look human, and makes publicly accessible a powerful new tool for future research on robots' human-likeness.

CCS CONCEPTS

• **Applied computing → Psychology**;

KEYWORDS

robot database, anthropomorphic robots, human-likeness, social robots

ACM Reference Format:
Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 2018. What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In *HRI '18: 2018 ACM/IEEE International Conference on Human-Robot Interaction, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3171221.3171268

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HRI '18, March 5–8, 2018, Chicago, IL, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-1-4503-4953-6/18/03…$15.00
https://doi.org/10.1145/3171221.3171268

1 INTRODUCTION

Robots are moving out of the realm of dull, dirty, and dangerous tasks and are increasingly taking on social tasks in everyday life [34]. They are beginning to interact with people in homes, workplaces, and other public spaces. Because many of the people interacting with these robots will be novice users with few prior experiences, their initial impressions are likely to be intuitively formed based on the robot's appearance.

Many current social robots have human appearance features such as eyes, hands, or faces. Although considerable research has investigated people's responses to human-like robots [16, 21, 29], there is very little systematic understanding of what makes a robot seem human-like. In the current paper, we introduce a tool for the systematic study of robots' human-like appearance—the ABOT (Anthropomorphic roBOT) Database. ABOT provides researchers and designers with images of 200 real-world anthropomorphic robots built for research or commercial purposes, and it accumulates data on people's perceptions of this wide variety of robots.

In addition, this paper presents two empirical studies that used the ABOT Database to examine what constitutes people's perception of a robot's human-likeness. Specifically, in the first study, we surveyed laypeople's (N = 1,140) judgments of whether various appearance features were present in each of the 200 robots in the ABOT database. A Principal Components Analysis (PCA) uncovered four distinct appearance dimensions (i.e., bundles of features) of human-like robots: (1) Surface Look (eyelashes, head hair, skin, genderedness, nose, eyebrows, apparel), (2) Body-Manipulators (hands, arms, torso, fingers, legs), (3) Facial Features (face, eyes, head, mouth), and (4) Mechanical Locomotion (wheels, treads/tracks). In the second study, we surveyed another 100 participants to assess their perceptions of each robot's overall physical human-likeness. Using multiple regression analyses, we were able to predict robots' perceived human-likeness from the presence of specific appearance features.

On the basis of these data we developed a tool available on the ABOT website that predicts how human-like a new robot will be perceived as a function of its specific appearance features.

2 BACKGROUND

2.1 Anthropomorphic Appearance as a Cornerstone of HRI Research

The psychological effects of robot appearance on human perceivers is a widely studied research topic in the field of human-robot interaction (HRI). Numerous studies have demonstrated that a robot's

appearance may considerably influence people's perceptions of its intelligence [18, 31], sociability [28], likability [7, 22], credibility [4], and submissiveness [35], among other characteristics and traits.

Even though robots, in theory, can take on any physical form, those modeled after humans have been shown to be particularly influential. For example, research by Krach et al. [20] reported that people experienced more fun interacting with increasingly human-like robotic partners, and Broadbent et al. [3] reported that users preferred a healthcare robot that displayed a human face over one without a human face. Other studies have similarly found that as a robot appears more human-like, people are more likely to take advice from the robot [28], take the robot's visual perspective [38], empathize with the robot [29], and even expect the robot to make moral decisions that are similar to those made by humans [21]. Such reactions may result from people attributing more mind to human-like robots [1, 20], as well as inferring positive characteristics such as being alive, sociable, and amiable [3]. Other researchers have argued that robots that resemble humans provide people with a sense of familiarity [16, 24], which may ease social acceptance [12].

At the same time, researchers have warned of certain risks associated with robots' human-like appearance. To start with, highly anthropomorphic robots may lead people to form expectations about capabilities that robots might not fulfill [1, 12, 26]. When those expectations are violated, people may lower their assessments of the robot [25], discontinue relying on the robot [19], and, in some cases, stop interacting with the robot altogether [9]. Furthermore, people may have aversive affective responses to highly human-like robots. For example, the "Uncanny Valley" hypothesis suggests that a robot's imperfect human-likeness can evoke eerie feelings in human perceivers [5, 22, 24, 32]. Other researchers have proposed that robots with a high degree of human-likeness may blur the human-robot boundary and undermine human uniqueness, leading people to perceive robots as a threat [15].

In sum, human-likeness in robots is undoubtedly powerful, yet it may not be universally desirable and requires careful assessment and consideration. Thus, it is necessary to develop a systematic understanding of robots with human-like appearance. For example, what are the features that make up anthropomorphic robots? Can these features be organized into a smaller number of underlying dimensions? And which features most strongly predict robots' perceived degree of human-likeness? Exploring these questions will fill important knowledge gaps and provide insight into the nuances of anthropomorphism in robots.

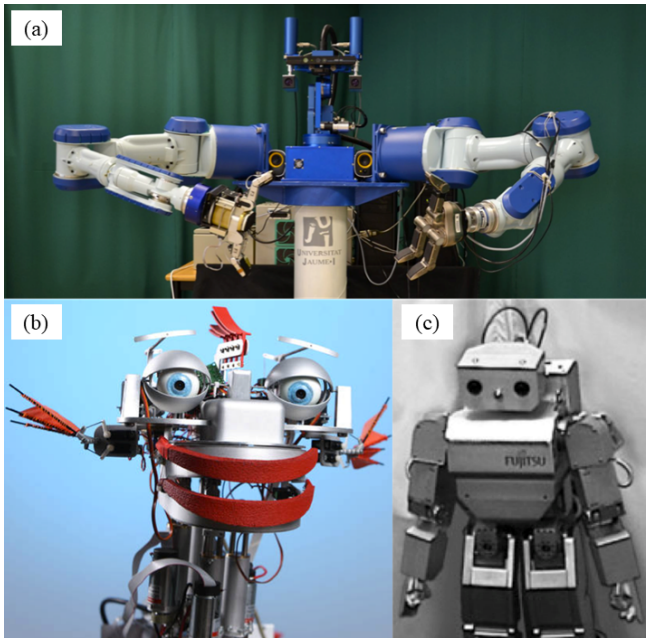
2.2 The Need for a Systematic Approach and Standardized Measures

As the number of anthropomorphic robots in research labs and on the commercial market steadily increases, the psychological effects of human-like robot appearance has drawn considerable attention from the research community and piqued public interest. However, exactly what constitutes human-likeness in robots remains vague and underspecified. There is currently no systematic, evidence-based approach for mapping robots on a continuum of perceived human-likeness. Consequently, researchers and designers are typically forced to rely on heuristics and intuitions when selecting human-like robots to include in studies or manipulating human-like features in robot design. This approach has suffered from several problems.

First, because a quantitative system to describe the degree of human-likeness in different robots is currently lacking, comparing research findings across studies has been challenging. So far, the most common practice has been to group robots into several broad categories based on their appearances, such as mechanical, humanoid, or android (e.g., [15, 20, 29, 30]). However, robots that share the same label across different studies may actually differ dramatically in their degree of human-likeness. For instance, compare the robots—all characterized as "humanoid"—in the following studies [23, 32, 35] (see Figure 1). Each of them was chosen as the "prototype" humanoid robot in their respective studies, yet it is unclear whether human perceivers would actually consider them to be equally human-like. A precise metric is therefore needed to compare different robots on a common scale and allow researchers to replicate findings with robots of equivalent human-likeness.

Second, even when researchers assess people's impressions of a robot's appearance quantitatively, they typically treat the concept of "human-likeness" as unidimensional. However, as illustrated by the three humanoid robots in Figure 1, human-likeness may present itself through different features. The robot in Stenzel et al. [33] has a head with exposed mechanics and no specific face contour, yet it has arms and hands. In contrast, the robot in Wiese et al. [36] consists of only a head with detailed human-like eyes, eyebrows, and a prominent mouth, yet exposed actuators without a smooth contour. Still different, the robot in Meltzoff et al. [23] has a clearly defined box-shaped head, a face with dark circles representing its eyes, a pointed structure representing its nose, two arms, and two hands with five fingers each. Each of these robots has been shown to create distinct psychological effects on human perceivers, but

Figure 1: Robots characterized as "humanoid" in (a) Stenzel et al. [33], (b) Wiese et al. [36], and (c) Meltzoff et al. [23].



we cannot tell whether these effects are all due to a broad human-likeness quality or due to specific features that make up a particular robot's human-likeness.

Third, when studying the effects of robot appearance, researchers have tended to limit themselves to robots that are either commercially successful (e.g., Nao, Robovie) or conveniently available to specific research groups (e.g., Kismet, Pearl). Although relying on only a few robots is likely a result of practical limitations, these practices could lead researchers to ignore the wide variety of extant robots and the nuanced differences in their appearances. Conclusions about the psychological effects of human-likeness may therefore not be generalizable to a more diverse collection of anthropomorphic robots.

By identifying the broad spectrum and specific features of robots' human-like appearance, a few studies have made progress in overcoming some of these problems. For instance, work by DiSalvo et al. [11] examined how features of robotic heads contributed to their perceived human-likeness. However, this work was limited to robot heads and included only 48 low-resolution images that came from research labs, commercial companies, or fictional depictions of robots in popular media. Mathur et al. [22] used 80 images of robots, but the images were also restricted to heads, and the representativeness of the image selection is unknown. A study by von der Pütten and Krämer [35] collected standardized full-body images of 40 anthropomorphic robots but focused on clustering robots by their familiarity and likeability, rather than by the presence of human-like appearance features. In addition, past projects have not made their robot images and corresponding ratings publicly available, thus limiting opportunities to conduct further research based on these data.

3 THE ABOT DATABASE

To address the above limitations and pave the way for more systematic, generalizable, and reproducible research on robots' human-like appearance, we created the ABOT (Anthropomorphic roBOT) Database and corresponding website. This project offers three benefits. First, it provides a survey of the broad landscape of anthropomorphic robots—indeed, the largest repository of robots with human-like features to date. Second, the ABOT Database provides standardized images of robots and an expanding dataset of people's perceptions of these robots, both publicly accessible for future research. Third, research using the ABOT Database will help deepen our understanding of what makes a robot look human. As a first step in this direction, we report two empirical studies that identify distinct dimensions of robot appearance and illuminate which of these dimensions ground people's perception of robots' overall human-likeness. In addition, we offer the Human-Likeness Estimator—a web-based linear equation that predicts how human-like a new robot will be perceived, based on the judged presence of a diagnostic set of appearance features.

We provide the aforementioned images, datasets, and tool on our ABOT Database website (http://www.abotdatabase.info/). The development of the database is described next.

3.1 Inclusion of Robots

In the first step to create the most extensive database of images of extant human-like robots, we set out to gather as many robots as possible that varied in both number and type of human-like appearance features. These robots were identified from multiple sources, including academic outlets (e.g., journal publications, conference proceedings), technology-focused media (e.g., online magazines, newsletters, blogs), websites of robotics companies and university laboratories, online robotics communities and discussion forums, and general Google searches. In addition, to identify robots that had not previously received wide media attention, we reached out to robotics researchers via robotics community listserves (e.g., HRI-ACM, Robotics Worldwide) to solicit images of robots. From this search and solicitation process between January 2017 and April 2017, we generated an initial image collection of 269 real-world robots with one or more plausible human-like features.

Next, we reviewed the collection of robots and excluded those whose appearance did not notably differ between versions (e.g., Cosero, Dynamaid) or had already been represented by other highly similar robots (e.g., several robots developed for soccer competitions). We also excluded robots that closely resembled animals. Then we reviewed the robot images and tried to ensure that at least one photograph of each robot could meet the following image standards: had no visual obstruction or motion blur, presented no other robots or humans, depicted the robot's whole body (except for android robots that could only be photographed sitting down), and was in color. When no satisfying images of a particular robot could be found, we often reached out to robot developers directly for their assistance. A robot was excluded only when all our attempts to obtain a photograph meeting our standards had failed. Using this procedure, we reduced our collection to 200 robots.

3.2 Image Selection and Editing

At this stage, most robots were represented in multiple images depicted from different angles and in various contexts. As a result, we identified the best image of each robot considering both image resolution and whether the robot body was presented in a standing, neutral, forward-facing posture (whenever available) with a neutral or mildly positive facial expression. Because many photographs were taken against cluttered backgrounds, our research team then edited those images using Adobe Photoshop to show all robots against transparent, white, or light-colored backgrounds.

4 STUDY 1: PRESENCE OF HUMAN-LIKE FEATURES AND HIGH-LEVEL DIMENSIONS

Study 1 took a bottom-up, feature-based approach to analyzing robot appearance. We aimed to document the human-like features present across anthropomorphic robots and use these data to identify underlying relationships among the features—that is, bundles of features that tend to co-occur in anthropomorphic robots. We expected that these bundles would reveal a smaller number of underlying dimensions that characterize the physical appearance of human-like robots. Rather than relying on experts' judgments, we asked laypeople to judge the presence or absence of 19 appearance features across robots in the ABOT Database.

4.1 Method

4.1.1 Participants. 1,140 participants were recruited for Study 1 via Amazon's Mechanical Turk crowdsourcing website (mTurk). Due to technical issues, data from 8 participants were not recorded. This resulted in data recorded from 1,132 participants (501 males, 619 females, 12 unreported) with ages ranging from 18 to 81 (M = 36.07, SD = 11.68, 4 unreported). Participants received $0.50 in compensation for the 5-minute study, which was approved by Brown University's Institutional Review Board (IRB).

4.1.2 Selecting and defining appearance features. To identify appearance features common across our database of robots, we developed a collection of possible robot appearance features (both human-like and non-human-like). We drew on a related coding scheme developed in [14], but we removed several categories such as "assumed purpose," ending up with 29 features in our initial feature collection.

In a pilot study, two independent raters (undergraduate research assistants) judged all robot images for the presence or absence of each feature in the collection. In light of their judgments, we eliminated 10 features that were rare, redundant, or too equivocal to be reliably distinguished (e.g., wings, digital vs. physical faces, eyelids, ears), resulting in a collection of 19 features to be used in the first study. In addition to human-like features such as eyes, head, and hands, the non-human-like features of wheels and treads/tracks were also included because locomotion is an important source of variation in people's imagined representations of robots [27]. Further, raters in the pilot study sometimes differed in how they interpreted human-like features in robots, so we decided to offer definitions of each feature to participants in the primary study. We began with relevant definitions for the features from the Oxford English and Merriam Webster's dictionaries. In adapting these definitions, we focused on retaining the human instantiation of the feature while removing descriptions of biological functions inapplicable to robots. For instance, the definition of mouth did not emphasize the intake of food or production of speech (see Table 1).

4.1.3 Design. Each of the 200 images of robots was randomly assigned to one of four blocks, consisting of 50 images each. When participants entered the study, they were randomly assigned to one of these four blocks and to one of the 19 appearance features (e.g., eyes); thus, they judged whether the given feature was present for each of the 50 robots in the assigned block. In addition, each participant also judged whether that same feature was present in eight images of humans (varying in age, gender, and ethnicity) and eight images of featureless smart home devices (e.g., the Amazon Echo, the Google Home, the Apple HomePod). These 16 additional images were used to establish clear anchors and to detect possibly confused or careless raters (e.g., claiming that images of humans did not have eyes or that featureless devices had arms). In total, each participant rated 66 images of robots, humans, and featureless devices on one of the 19 appearance features, with the order of presentation of images randomized for each participant. We planned to assign approximately fifteen judges to each robot in each block on each feature, requiring a total sample of N = 1,140 (15 raters x 19 features x 4 blocks of robots).

Table 1: Collection of appearance features and associated definitions in Study 1.

Feature:	Definition
Apparel:	Materials worn temporarily to cover the body.
Arm:	Upper limb typically used for manipulating objects.
Eye:	A round or oval shaped form that often gathers visual information.
Eyebrow:	A line above the eye usually consisting of hair.
Eyelashes:	Threadlike filaments that surround the eyelid.
Face:	The front part of the head, which may contain features such as eyes, nose, or mouth.
Finger:	Each of a number of slender jointed parts connected to the hand.
Genderedness:	Features of appearance that can indicate biological sex, or the social categories of being male or female.
Gripper:	The claw-like terminal part of an appendage used for grasping and manipulating objects. (A claw-like gripper is not a hand).
Hand:	The terminal part of an arm, typically connected to the arm by a wrist. A hand is normally used for grasping, manipulating, or gesturing. (A claw-like gripper is not a hand).
Head:	The uppermost part of a body, typically connected to the torso by a neck. The head may contain facial features such as the mouth, eyes, or nose.
Head hair:	A collection of threadlike filaments on the head.
Leg:	Lower limb used for movement over ground.
Mouth:	A large opening located on the lower part of the face.
Nose:	A projected feature of the face above the mouth.
Skin:	A thin layer of tissue covering almost the entire body.
Torso:	The trunk or middle part of the body (minus the limbs and head).
Treads/Tracks:	Moving bands that transport things like tanks over rough terrain. They are sometimes called tank treads or caterpillar treads.
Wheel:	A round device that rolls on the ground and transports an entity over surfaces.

4.1.4 Procedure. After reviewing the informed consent form, participants learned that the study investigated how people perceive the appearance of robots and other entities. They were provided with the definition of one of the features as shown in Table 1, and they judged whether the entities in each of the 66 images possessed that feature by clicking "yes" or "no."

Participants were given unlimited time to judge whether or not the given feature was present in each of the images. Halfway through the task, participants received a short (approximately 10-second) break. During the break, participants were reminded of the relevant feature definition. At the end of the study, participants completed a brief demographic questionnaire about age, gender, and native language, and allowed participants to provide feedback about the study. Once complete, participants received a payment code. The entire study took approximately 5 minutes to complete.

4.2 Results

4.2.1 Data screening and rater reliability. We first identified participants who warranted exclusion by at least one of two criteria: they marked three or more of the 16 "catch trials" (additional images of humans and featureless devices) in the obviously incorrect way; or their profile of responses substantially deviated from the group average. In analogy to item-total correlations (r_{iT}) in psychometric test analyses, we computed correlations between each rater's judgments (across a block of 50 robots) with the average of the remaining raters' judgments in the same block and for the same feature, r_{rG}. Just as $r_{iT} < .30$ typically marks items that measure something other than the scale as a whole and should be eliminated from the scale [10], we set $r_{rG} < .30$ to identify raters that did not perform the same judgment task as the group as a whole and should likewise be eliminated from remaining analyses. Using these two criteria, we eliminated 73 out of 1132 raters. All reported analyses are based on data obtained from the remaining 1,059 raters. Technical issues caused data loss for 2 of the robots, so these 1,059 raters judged 198 robots in total.

Next we computed inter-rater reliabilities, using a variant of Cohen's *kappa* (κ) [8]. The original κ was intended to correct for chance agreement among raters, but it is known to be misleading when marginal response distributions are extreme [6]. Several of our data vectors have such extreme distributions (e.g., when almost all participants agree that most robots have a torso or that very few robots have eyebrows). Several corrected κ computations have been proposed, and we selected Gwet's AC1 measure, derived mathematically and supported by Monte Carlo simulations [17, 37]. The inter-rater reliabilities for the 19 features were generally high, average AC1 = 0.71, ranging from 0.48 (treads) to 0.89 (headhair), with only three below 0.60. One feature, gripper, was problematic, both because of its lower reliability for robots (AC1 = 0.50) and because 22 out of 60 raters indicated that human images displayed a "gripper." Indeed, numerous participants expressed confusion over the distinction between grippers and hands or fingers. For these reasons, we excluded the gripper feature from further analyses.

4.2.2 Feature-present scores. We computed *feature-present scores* by calculating the proportion of raters who endorsed the presence of a given feature for a given robot. For example, if 13 out of 15 raters indicated the presence of the feature "head" on a given robot, this robot's feature-present score for "head" would be 0.87. Thus, each robot had a score between 0 (no raters indicated the feature was present) and 1 (all raters indicated the feature was present) for each of the 18 analyzed features. Scores closer to 1 indicate high consensus that the given robot had the particular feature, while scores closer to 0 indicate high consensus that the given robot did not have that feature; scores closer to 0.5 indicate that raters disagreed over whether the given robot had that feature.

Detailed documentation of each robot's feature-present scores can be downloaded from our ABOT Database website.

Table 2: Principal Components Loading Matrix, Study 1.

Feature	PC 1 Surface	PC 2 Body	PC 3 Facial	PC 4 Mech.
1. Eyelashes	**.88**	-.08	.15	-.04
2. Headhair	**.85**	.05	.03	-.06
3. Skin	**.83**	.07	.07	-.13
4. Genderedness	**.80**	.28	.17	-.12
5. Nose	**.71**	.05	.33	-.05
6. Eyebrows	**.69**	-.19	.38	.02
7. Apparel	**.68**	.28	.07	-.13
8. Hands	.12	**.93**	.06	.02
9. Arms	.02	**.92**	.10	.01
10. Torso	.07	**.90**	.19	.07
11. Fingers	.14	**.86**	.02	.05
12. Legs	-.06	**.74**	-.08	-.23
13. Face	.28	.14	**.90**	.02
14. Eyes	.14	-.02	**.88**	-.01
15. Head	.13	.49	**.73**	.03
16. Mouth	.48	.05	**.57**	-.07
17. Wheels	-.13	-.09	.01	**.92**
18. Treads/Tracks	-.18	.06	-.01	**.91**
Eigenvalue	4.67	4.30	2.81	1.79
% Variance	25.93	23.88	15.62	9.93
Subscale Cronbach's α	.89	.93	.83	.82

Note: **PC 1**: Surface Look, **PC 2**: Body-Manipulators, **PC 3**: Facial Features, **PC 4**: Mechanical Locomotion. Subscales derived from features with loadings in bold.

4.2.3 Discovery of appearance dimensions. We examined whether the 18 remaining appearance features were organized into correlated bundles by conducting a Principal Components Analysis (PCA) on the feature-present scores across 198 robots. The PCA yielded four high-level dimensions (principal components) that had eigenvalues greater than 1 and together explained 75.36% of the total variance. After Varimax rotation, Component 1 revealed strong loadings (> .50) for seven surface features that are most commonly seen in androids (eyelashes, head hair, skin, genderedness, nose, eyebrows, and apparel). Component 2 revealed strong loadings for five features characterizing a human-like body and manipulators (hands, arms, torso, fingers, and legs). Component 3 revealed strong loadings for four features characterizing a human-like face (face, eyes, head, and mouth). Finally, Component 4 revealed strong loadings for two features characterizing a robot's mechanical locomotion (wheels and treads/tracks). These components and the loadings of each of the features are shown in Table 2.

4.2.4 Component scores and creation of subscales. The PCA in Study 1 yielded four appearance dimensions (i.e., feature bundles), which we labeled Surface Look, Body-Manipulators, Facial Features, and Mechanical Locomotion. To locate each robot in this four-dimensional space, we computed two measures. The first consists of the four "principal component scores" calculated using the regression method in SPSS version 24 (see [13]). Each component

score reflects a weighted linear combination of all 18 feature scores, with precise weights (factor coefficients) assuring that the components are uncorrelated. A second, more interpretable measure is the "subscale scores," which we computed by averaging feature-present scores of those features that loaded highly on each respective dimension (the bold-faced items in Table 2). Because feature-present scores range from 0 to 1, the resulting four subscale scores also range from 0 (none of the defining features for this dimension were present) to 1 (all defining features were present).

Overall, the subscale scores intuitively describe how strongly a robot embodies each of the four appearance dimensions. To evaluate how well they correlated with the regression-based component scores, we calculated Pearson's correlation coefficients between them for all four dimensions. The results confirmed that the subscales captured the original four principal components very well (Surface Look, $r = .96$, $p < .001$; Body-Manipulators, $r = .99$, $p < .001$; Facial Features, $r = .92$, $p < .001$; Mechanical Locomotion, $r = .98$, $p < .001$). Consequently, the easily calculable and interpretable subscale scores are practically equivalent to component scores derived from the entire set of features. Thus, we recommend calculating and reporting them as an efficient method to summarize any given robot's appearance on four distinct dimensions.

4.3 Discussion

In Study 1, we surveyed the broad landscape of anthropomorphic robots contained in the ABOT Database and documented people's judgments of the presence of appearance features in each robot. Based on systematic correlations among these appearance features, we were able to identify four dimensions of human-like robot appearance: (1) Surface Look, (2) Body-Manipulators, (3) Facial Features, and (4) Mechanical Locomotion. Together, these four dimensions accounted for three-fourths of the total variance among the 18 individual features. Selecting the highest-loading features of each component, we computed parsimonious subscale scores that mark each robot's standing on the four dimensions and thus offer a precise and quantitative profile of a robot's human-like appearance.

Our next goal was to demonstrate the psychological relevance of the appearance dimensions by showing how they constitute overall impressions of human-likeness.

5 STUDY 2: PREDICTING PHYSICAL HUMAN-LIKENESS

In Study 2, we assessed people's overall judgments of robots' human-likeness and examined how well these judgments are predicted by the four appearance dimensions discovered in Study 1 and by specific appearance features. We thus identified which appearance dimensions best characterize general human-likeness impressions, bringing more clarity to the vague concept of human-likeness. In addition, we created a tool that allows researchers to predict how human-like a new robot would be perceived by laypeople based on the presence of appearance features.

5.1 Method

5.1.1 Participants. 100 participants entered into this study, but data for two participants were lost due to Internet connection issues. Thus, we analyzed data from 98 participants (48 males, 50 females) with ages ranging from 19 to 64 ($M = 33.42$, $SD = 9.75$). Participants completed the study via mTurk in exchange for $1.00 in compensation.

5.1.2 Overall human-likeness measure. For each robot, approximately 25 participants were assigned to judge how physically human-like the robot appeared overall. Participants provided their judgment by dragging a slider to indicate where each robot fell on a continuum between "Not human-like at all" (left-most point of the slider's scale) to "Just like a human" (right-most point of the slider's scale). The chosen slider position was then converted into a number ranging from 0 (Not human-like at all) to 100 (Just like a human).

5.1.3 Design. As in Study 1, participants were randomly assigned to one of the four blocks of 50 images of robots, along with images of eight humans and eight featureless smart home devices. In total, participants rated images of 66 entities for "how much each entity in the photograph looks like a human in its physical appearance." We planned to assign 25 judges to each robot in each block and thus predetermined a total sample of $N = 100$.

5.1.4 Procedure. After reviewing the informed consent form, participants read instructions on how to use the slider scale to judge the human-like physical appearance of the 66 entities. They were given as much time as needed to make each judgment. Halfway through rating the images, participants received a 10-second break. At the end of the study, participants were asked to fill out a brief demographic questionnaire identical to that in Study 1. Once complete, participants were provided with an mTurk confirmation code to receive their payment. The entire study took approximately 5 minutes to complete.

5.2 Results

5.2.1 Data screening and rater reliability. In data screening, only 2 of the 98 raters had to be excluded. One gave arguably incorrect responses for most featureless machines (human-likeness scores of 30 or more) and one provided uniform responses (slider at 1) to nearly all robots. Inter-rater reliability of human-likeness scores was assessed via the intra-class correlation ICC(2,1), which averaged 0.60 across the four blocks. Further, correlations between each rater's judgments and the average of the remaining raters' judgments, r_{iG}, averaged .81 across the four blocks. Thus, the group averages, used in subsequent analyses, were highly representative of any individual raters' judgments.

5.2.2 Human-likeness scores. We then averaged the scores across raters to provide an overall physical human-likeness score for each robot. These average human-likeness scores across all the robots in our database ranged from 1.44 to 96.46, with $M = 33.26$, $SD = 18.97$. This distribution confirms that the ABOT database includes a wide spectrum of human-like robots.

5.2.3 Predicting physical human-likeness from feature-present scores. In order to predict robots' human-likeness with a feature-based approach, we used the robots' 18 individual feature-present scores as predictors of their overall human-likeness scores. A regression model using all 18 features explained 88.8% of the total variance of overall human-like scores ($R = .94$, $F(18, 179) = 78.5$, $p <$

.001). Given the impressive predictive power of this linear model, we built a *Human-Likeness Estimator* from it, which can be accessed from our ABOT Database website. This estimator serves to approximate how human-like a robot will be perceived by laypeople based on the presence or absence of its appearance features.

Finally, in order to identify a small number of physical features that as a minimal group most effectively predict the perceived human-likeness of a robot, we conducted a stepwise forward regression analysis with the 18 features as predictors. The top three predictors were torso ($r_{semi-partial}$ = .31), genderedness (r_{sp} = .44), and skin (r_{sp} = .23), all ps < .001, together explaining 78.4% of the variance of overall human-likeness judgments, $F(3, 194)$ = 234.06, $p < .001$.

5.2.4 Predicting physical human-likeness from appearance dimensions. Next we conducted two multiple regression analyses to examine how well people's overall human-likeness judgments can be predicted from the four high-level appearance dimensions discovered in Study 1.

The first analysis used the four regression-based principal component scores as predictor variables. This model explained 82.5% of the variance in human-likeness, $F(4, 193)$ = 227.0, $p < .001$. The strongest predictors were Surface Look (with 37.2% explained variance, r_{sp} = .61) and Body-Manipulators (with 36.0% explained variance, r_{sp} = .60), but each of the other components also added unique explained variance: 5.7% for Facial Features (r_{sp} = .24), and 3.6% for Mechanical Locomotion (r_{sp} = -.19) (all ps < .001). Further, cross validation bootstrapping procedures revealed that the unique explained variances in the overall model and those in bootstrapped samples of the model only differed from one another by an average of 1.5% for Surface Look, 3.4% for Body-Manipulators, 0.7% for Facial Features, and 0.3% for Mechanical Locomotion.

The second analysis used the four subscale scores as predictors. This model explained 81.5% of the variance in human-likeness, $F(4, 193)$ = 212.4, $p < .001$. The strongest predictor was the Body-Manipulators dimension with 28% of variance explained (r_{sp} = .53, $p < .001$). Other dimensions also added unique explained variance: Surface Look explained 19% variance (r_{sp} = .44, $p < .001$), Mechanical Locomotion explained 1.7% (r_{sp} = -.13, $p < .001$), and Facial Features explained 0.5% (r_{sp} = .07, p = .025). Similar to above, the difference between the explained variance in the bootstrap model and the full model was small: average difference in unique explained variance was 2.8% for Body-Manipulators, 0.5% for Surface Look, 0.3% for Facial Features and 0.1% for Mechanical Locomotion. More information about our cross validation estimates can be found on the ABOT Database website.

5.3 Discussion

In Study 2, we surveyed laypeople's perceptions of robots' overall physical human-likeness and showed that this overall judgment can be strongly predicted from either specific individual appearance features (especially torso, genderedness, and skin) or the underlying four appearance dimensions (especially Body-Manipulators and Surface Look). At a theoretical level, the strength of this prediction suggests that the broad human-likeness concept can be decomposed into meaningful appearance dimensions. At the practical level, the strength of this prediction allowed us to build a Human-likeness Estimator that can be used to quickly estimate how human-like a new robot will be perceived.

6 GENERAL DISCUSSION

The psychological effects of robots' appearance, and especially human-like appearance, form a cornerstone of HRI research. Despite the topic's broad appeal in scholarship and public perception, a systematic understanding of what human-likeness in robots really is has been lacking. Our research has aimed to illuminate the concept of human-likeness by taking a feature-based approach—decomposing the broad concept into specific appearance features and identifying underlying high-level dimensions. And by compiling all the images and data we collected into a publicly accessible database, we hope to pave the way for more systematic, generalizable, and reproducible research on robots' anthropomorphic appearance and its effects on human-robot interaction.

6.1 Theoretical Contributions

In Study 1, we asked laypeople to judge the presence of human-like appearance features in 200 robots, and nearly all these features were assessed with good inter-rater reliability. More importantly, the features co-occurred in robots in systematic ways and revealed four underlying appearance dimensions: Surface Look (e.g., gender, skin, eyelashes) Body-Manipulators (e.g., torso, arms, hands), Facial Features (e.g., head, eyes), and Mechanical Locomotion (i.e., wheels, treads/tracks). These dimensions are not the result of people's *mere assumptions* about the features that co-occur in robots, because people never made judgments about more than one feature at a time. Instead, the dimensions reflect underlying co-occurrences of features in the actual robots that have been built for research or commercial purposes. Thus, with just four subscale scores (averaged feature-present scores on the four appearance dimensions) we can describe the distinct profiles of hundreds of extant anthropomorphic robots.

In Study 2, we showed that robots' appearance profiles obtained in Study 1 could predict their perceived overall physical human-likeness. The strong predictive power of the four appearance dimensions (more than 80% explained variance) reveals that people's impression of a robot's human-likeness is a result of variations in a few fundamental dimensions of appearance. In particular, the dimensions of Surface Look and Body-Manipulator strongly contributed to robots' perceived human-likeness, and facial features and (the absence of) mechanical forms of locomotion made weaker but still notable contributions.

The powerful impact of these appearance dimensions on overall human-likeness perceptions is unlikely to be a mere physical coincidence; rather, it conveys important psychological expectations people have of robots—and human-like robots in particular. The Body-Manipulators dimension reflects people's expectations that robots interact with the physical environment in the same manner as humans—and because many of the objects in the physical world have been designed for human biological manipulators, robots may need those manipulators as well. The dimension of Facial Features reflects people's expectations that robots socially interact and effectively communicate with humans. That is, even though human-like heads, eyes, and mouths on robots do not serve biological functions

as with humans (e.g., seeing, eating, speaking), they serve critical communicative functions such as facilitating joint attention via gaze orientation [23] or providing non-verbal responses by nodding and shaking one's head [2]. Finally, features in the Surface Look dimension (e.g., skin, apparel, genderedness) are less relevant to a robot's task-specific capacities but do mimic the human veneer, engendering familiarity, comfort, and perhaps trustworthiness [22].

6.2 Practical Contributions

With the ABOT Database, we are making publicly available a repository of standardized, high-quality images of anthropomorphic robots, which surveys the wide variety of anthropomorphic robots extant in the world today. Each robot now has an appearance profile, summarized by their appearance feature scores and dimension scores, and we intend to create such profiles for new robots that will be added to the database. Moreover, each robot has an overall human-likeness score, and we intend to collect other psychological characteristics associated with each robot in the near future.

Using the ABOT Database, researchers can compare different robots across previous studies and draw more detailed conclusions about the psychological effects of specific aspects of human-likeness. Moreover, using the Human-Likeness Estimator, researchers and designers can effectively describe new or planned robots on important appearance features and make reasonable predictions about laypeople's perceptions of those robots' overall human-likeness. In addition, our research may offer a general recipe to build robots that do or do not appear human-like. For example, to create robots that should appear human-like, designers may focus on adding more features in the Surface Look dimension (gender, skin, eyelashes); conversely, to reduce human-likeness, designers may best install treads or wheels rather than legs.

6.3 Limitations and Future Directions

There are several limitations to the current research. First, the inclusion of robots in the ABOT Database is constrained by both our knowledge about extant robots and our search procedures. As a result, despite our efforts to create the most comprehensive anthropomorphic robot collection, we are aware that the database presents only a subset of anthropomorphic robots ever created. Furthermore, new robots are continuously being developed in research laboratories and released to the commercial markets, and their designs are becoming increasingly diverse. Therefore, we invite other researchers and designers to inform us of the robots that are currently missing in this database, and we plan to update and re-run analyses as the database grows.

In addition, although we identified a considerable number of appearance features that constitute four high-level dimensions and predict overall human-likeness, we think the number and kinds of features to describe robots may be refined. For example, we found it difficult to identify ears, lips, or eyelids on the full range of current human-like robots. We cannot rule out that such features might influence overall perceptions of human-likeness, nor can we anticipate their impact on perceptions of other robot characteristics. Perhaps by subdividing the robot sample or working with more detailed 360° images or videos, even the most subtle features may become available for research.

Second, our analyses of the current dataset are straightforward, but additional findings may be uncovered with advanced statistical analyses. Thus, we encourage other researchers to use the ABOT Database to probe the nuances of human-likeness in robots. For example, while the PCA ensured that the four principal components are linearly independent, the subscales are somewhat correlated and may even be correlated non-linearly, indicating additional important relationships among these dimensions. As another example, several feature-present scores are non-independent, as the presence of some features (e.g., eyebrows) presupposes other features (e.g., eyes). Analyzing conditional probabilities among features will likely reveal a hierarchical structure of various appearance features and further refine our current decomposition of human-likeness.

Third, we asked participants only to diagnose whether certain classes of appearance features were present or absent in robots; our research did not take into account variations in the appearance of specific features. For instance, one can imagine that the same feature (e.g., eyes) could be represented in different ways (e.g., biologically realistic, abstract, or cartoonish) [12]. It is possible that variations in these features may also influence perceptions of human-likeness.

Finally, our assessment focused solely on investigating how physically human-like a robot might be perceived given static images. There are many others ways in which robots' human-likeness can be expressed, such as through dynamic movement, speech, nonverbal gestures, or emotional expressions. Therefore, how static appearance features of robots may interact with such dynamic characteristics is a meaningful future research direction.

7 CONCLUSION

The human-like appearance of robots is a centerpiece of research in the HRI domain. However, defining human-like appearance in robots has remained elusive. This research provides a systematic investigation of the human-like features present in anthropomorphic robots, uncovers high-level structures interconnecting those features, reveals their relationship to robots' overall physical human-likeness, and offers methodological tools for determining if a particular robot will be considered human-like in its appearance.

8 ACCESSING THE ABOT DATABASE

The ABOT Database is available at http://www.abotdatabase.info/

ACKNOWLEDGMENTS

The authors thank Salomi Aladia, Mckenna Cisler, Fue Vue, Broderick Allan, and Maya Menefee for their contributions to this project as well as all the researchers who responded to our email requests and contributed their robot images to our database. This work is supported by Office of Naval Research grant #N00014-14-1-0144 and by the Brown University Humanity-Centered Robotics Initiative. Daniel Ullman is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

REFERENCES

[1] Christoph Bartneck, Takayuki Kanda, Omar Mubin, and Abdullah Al Mahmud. 2009. Does the design of a robot influence its animacy and perceived intelligence? *International Journal of Social Robotics* 1, 2 (2009), 195–204.

[2] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*. IEEE, 708–713.

[3] Elizabeth Broadbent, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q Stafford, Bruce A MacDonald, and Daniel M Wegner. 2013. Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a better personality. *PloS One* 8, 8 (2013), e72589.

[4] Judee K Burgoon, Joseph A Bonito, Bjorn Bengtsson, Carl Cederberg, Magnus Lundeberg, and L Allspach. 2000. Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior* 16, 6 (2000), 553–574.

[5] Tyler J Burleigh, Jordan R Schoenherr, and Guy L Lacroix. 2013. Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior* 29, 3 (2013), 759–771.

[6] Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46, 5 (1993), 423–429.

[7] Álvaro Castro-González, Henny Admoni, and Brian Scassellati. 2016. Effects of form and motion on judgments of social robots' animacy, likability, trustworthiness and unpleasantness. *International Journal of Human-Computer Studies* 90 (2016), 27–38.

[8] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[9] Maartje MA de Graaf, Somaya Ben Allouch, and Jan AGM van Dijk. 2017. Long-term evaluation of a social robot in real homes. *Interaction Studies* 17, 3 (2017), 462–491.

[10] David De Vaus. 2002. *Analyzing social science data: 50 key problems in data analysis*. Sage.

[11] Carl F DiSalvo, Francine Gemperle, Jodi Forlizzi, and Sara Kiesler. 2002. All robots are not created equal: The design and perception of humanoid robot heads. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*. ACM, 321–326.

[12] Brian R Duffy. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42, 3 (2003), 177–190.

[13] Ryne Estabrook and Michael Neale. 2013. A comparison of factor score estimation methods in the presence of missing data: Reliability and an application to nicotine dependence. *Multivariate Behavioral Research* 48, 1 (2013), 1–27.

[14] Neta Ezer. 2008. *Is a robot an appliance, teammate, or friend? Age-related differences in expectations of and attitudes towards personal home-based robots*. Georgia Institute of Technology.

[15] Francesco Ferrari, Maria Paola Paladino, and Jolanda Jetten. 2016. Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics* 8, 2 (2016), 287–302.

[16] Julia Fink. 2012. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics*. Springer, 199–208.

[17] Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48.

[18] Kerstin S Haring, David Silvera-Tawil, Tomotaka Takahashi, Katsumi Watanabe, and Mari Velonaki. 2016. How people perceive different robot types: A direct comparison of an android, humanoid, and non-biomimetic robot. In *Knowledge and Smart Technology (KST), 2016 8th International Conference on*. IEEE, 265–270.

[19] Takanori Komatsu, Rie Kurosawa, and Seiji Yamada. 2012. How does the difference between users' expectations and perceptions about a robotic agent affect their behavior? *International Journal of Social Robotics* 4, 2 (2012), 109–116.

[20] Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. 2008. Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS One* 3, 7 (2008), e2597.

[21] Bertram F Malle, Matthias Scheutz, Jodi Forlizzi, and John Voiklis. 2016. Which robot am I thinking about?: The impact of action and appearance on people's evaluations of a moral robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 125–132.

[22] Maya B Mathur and David B Reichling. 2016. Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition* 146 (2016), 22–32.

[23] Andrew N Meltzoff, Rechele Brooks, Aaron P Shon, and Rajesh PN Rao. 2010. "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks* 23, 8 (2010), 966–972.

[24] Masahiro Mori. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.

[25] Steffi Paepcke and Leila Takayama. 2010. Judging a bot by its cover: An experiment on expectation setting for personal robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 45–52.

[26] Elizabeth Phillips, Scott Ososky, Janna Grove, and Florian Jentsch. 2011. From tools to teammates: Toward the development of appropriate mental models for intelligent robots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 55. SAGE Publications Sage CA: Los Angeles, CA, 1491–1495.

[27] Elizabeth Phillips, Daniel Ullman, Maartje de Graaf, and Bertram F. Malle. 2017. What does a robot look like?: A multisite examination of user expectations about robot appearance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA.

[28] Aaron Powers and Sara Kiesler. 2006. The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. ACM, 218–225.

[29] Laurel D Riek, Tal-Chen Rabinowitch, Bhismadev Chakrabarti, and Peter Robinson. 2009. Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 1–6.

[30] Michihiro Shimada, Takashi Minato, Shoji Itakura, and Hiroshi Ishiguro. 2006. Evaluation of android using unconscious recognition. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*. IEEE, 157–162.

[31] Valerie K Sims, Matthew G Chin, David J Sushil, Daniel J Barber, Tatiana Ballion, Bryan R Clark, Keith A Garfield, Michael J Dolezal, Randall Shumaker, and Neal Finkelstein. 2005. Anthropomorphism of robotic forms: A response to affordances?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49. SAGE Publications Sage CA: Los Angeles, CA, 602–605.

[32] Jan-Philipp Stein and Peter Ohler. 2017. Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition* 160 (2017), 43–50.

[33] Anna Stenzel, Eris Chinellato, Maria A Tirado Bou, Ángel P del Pobil, Markus Lappe, and Roman Liepelt. 2012. When humanoid robots become human-like interaction partners: Corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance* 38, 5 (2012), 1073.

[34] Leila Takayama, Wendy Ju, and Clifford Nass. 2008. Beyond dirty, dangerous and dull: What everyday people think robots should do. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 25–32.

[35] Astrid Marieke von der Pütten and Nicole C Krämer. 2012. A survey on robot appearances. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 267–268.

[36] Eva Wiese, Agnieszka Wykowska, Jan Zwickel, and Hermann J Müller. 2012. I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PloS One* 7, 9 (2012), e45391.

[37] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology* 13, 1 (2013), 61.

[38] Xuan Zhao, Corey Cusimano, and Bertram F Malle. 2016. Do people spontaneously take a robot's visual perspective?. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE, 335–342.