A Comprehensive Approach to Validating the Uncanny Valley using the Anthropomorphic RoBOT (ABOT) Database

Boyoung Kim
Warfighter Effectiveness Research Center
United States Air Force Academy
Air Force Academy, Colorado
boyoung.kim.kr.ctr@usafa.edu

Micala Bruce
Dept. of Systems Engineering
United States Air Force Academy
Air Force Academy, Colorado
c20micala.bruce@usafa.edu

LeSean Brown
Dept. of Systems Engineering
United States Air Force Academy
Air Force Academy, Colorado
c20lesean.brown@usafa.edu

Ewart de Visser
Warfighter Effectiveness Research Center
United States Air Force Academy
Air Force Academy, Colorado
ewartdevisser@gmail.com

Elizabeth Phillips
Warfighter Effectiveness Research Center
United States Air Force Academy
Air Force Academy, Colorado
elizabeth.phillips@usafa.edu

***Abstract*—The uncanny valley hypothesis posits that people's emotional responses to robots are increasingly positive as robots' resemblance to humans increases. However, when robots closely, but imperfectly resemble humans, people's responses turn negative, only to revert back once their appearance more closely resembles humans. These sharp emotional transitions (i.e., peaks and valleys in emotional response) from positive to negative, and then back to positive, are collectively referred to as the uncanny valley. In this project, we attempted to validate the uncanny valley with the largest set of real-world robots currently available in open source format (the ABOT Database). Participants saw static images of 251 robots which varied in their degree of human-likeness, and rated them on uncanniness. We found significant empirical support not only for the hypothesized uncanny valley but an additional valley. This unanticipated valley emerged when the robots' appearance had low to moderate human-likeness. Unique combinations of appearance dimensions of human-like robots may be responsible for the presence of an additional valley for robots that only moderately resemble humans. These findings of uncanny valleys in the existing robots may have important implications for robot design.**

***Index Terms*—Uncanny valley, human-likeness, human-robot interaction, anthropomorphic, database**

I. Introduction

The uncanny valley [1, 2] refers to hypothesized changes in emotional responses toward robots as their physical resemblance to humans (i.e., human-likeness) increases. Specifically, the uncanny valley hypothesis posits that emotional responses to robots become increasingly positive as their appearance resembles that of humans; but when robots resemble humans too closely, people become disturbed, unnerved, uneasy, or perceive such robots as uncanny. And only after robots appear nearly indistinguishable from humans, do emotional responses become positive again (See Fig.1).

There have been many attempts to examine the effects of robot human-likeness on people's emotional responses to robots. But prior studies have used a relatively small range of robot exemplars or focused on specific parts of the robot's body [4, 5, 6, 7, 8, 9]. For instance, Rosenthal-Von Der Pütten and Krämer [10] used full-body images of robots, but the number of robot images was limited to 40. By contrast, Mathur et al. [5] investigated emotional responses to 80 real-world robots, but their focus was constrained to robot heads.

Although relying on a small set of robot exemplars is likely the result of practical limitations, this approach may overlook the theoretical and practical implications of considering the vast range and variety of the extant human-like robots when investigating the uncanny valley. For example, Phillips, Zhao, Ullman, and Malle examined 251 real-world human-like robots and found that human-like appearance in robots can be decomposed into three human-like appearance dimensions: the robots' Surface features (e.g., skin, hair, apparel), the

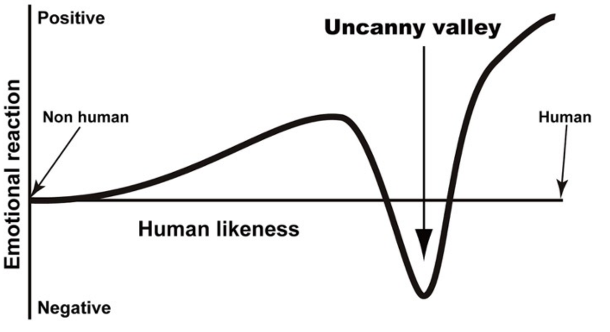

Fig. 1. Graphical representation of the Uncanny Valley. Retrieved from [3].

978-1-7281-7145-6/20/$31.00 ©2020 IEEE

main components of the robots' Body-manipulators (e.g., Torso, Arms, Legs), and the robots' Facial features (e.g., eyes, mouth, face) [11]. These findings suggest that people's overall perception of robots' physical human-likeness and its relationship with emotional responses to the robots may be explained by different constellations of the three human-like appearance dimensions. If the hypothesized uncanny valley phenomenon could be understood at the level of specific human-like appearance dimensions, it could lead to improved robot designs.

As a first step to better the understanding of the uncanny valley hypothesis [1], the present research sought to validate the existence of the uncanny valley using the largest collection of real-world robots to date. Specifically, we showed participants 251 full-body robot images available in the Anthromorphic roBOT(ABOT) database, and asked them to rate the robots on the perceived uncanniness. The robots in the ABOT database represent a wide sample of robots on the human-likeness spectrum—ranging from 1.44 to 96.38 on a scale of 0 to 100—empirically derived and made widely available to date [11]. Thus, if Mori's uncanny valley hypothesis held true for the real-world robots, the participants would increasingly rate the robots as uncanny as the appearance of the robots becomes highly, but not yet perfectly, human-like.

In addition to testing the uncanny valley, the current project explored how people emotionally react to robots that represent the vast majority of real-world human-like robots—approximately 82% of the existing robots in the database have human-likeness scores less than 50 on a 0-100 rating scale. In fact, the ABOT database [11] shows that the majority of real-world human-like robots fall in the lower third of the human-likeness spectrum ($M = 33.81$, $Median = 31.76$). Some of the robots within this range of the human-likeness spectrum have strange combinations of human-like features (e.g., the presence of some surface features like eyebrows, but missing many other features of the face). Prior researchers have argued that perceptual mismatches between appearance features can explain uncanny reactions to robots [12]. If this explanation for the uncanny valley hypothesis is valid, the "valley(s)" in emotional responses may also appear lower on the spectrum of human-likeness, where robots with unusual combinations of appearance dimensions reside. This suggests that even robots relatively low in human-likeness may be perceived as highly uncanny, which would be a prediction not captured by the original uncanny valley hypothesis [1, 2]. Thus, the purpose of this project was to investigate (1) whether the presence of the uncanny valley persists over the largest range of human-like robots to date and (2) explore whether there are additional valleys present for less human-like robots.

II. Method

A. Participants

To obtain ratings from 30 participants per robot, we grouped all robots into one of five blocks, then aimed to recruit 150 participants (30 per 5 blocks) to complete the study on Amazon's Mechanical Turk (MTurk). Data collected on crowd-sourcing platforms like Amazon's Mechanical Turk is becoming increasingly common, and empirical studies comparing the quality of MTurk data to traditional laboratory studies have shown the data to be comparable [13]. However, some researchers have noticed a recent decline in the quality of data acquired on MTurk [14]. Thus, to ensure that we only included reliable data in our analyses, we applied a rigorous four-step data screening process. First, we discarded data obtained from participants who provided ratings for less than half of the trials. Second, we eliminated data submitted by participants who answered incorrectly to six or more of 16 catch-trials included in the stimuli set. Third, we considered a lack of variation in intra-participant ratings as an indicator of inattentiveness. Hence, we removed data from participants whose ratings had a standard deviation less than 10 ($SD < 10$) across a scale of $0 - 100$. Lastly, we compared each participant's ratings to the average of the remaining rater judgments in their group (inter-participant), by computing the correlation between individual rater judgements to the remaining judgements in their group, r_{rG}. If this correlation between the individual participant's ratings and the group mean was smaller than 0.30 ($r_{rG} < .30$), we discarded the participant's data, as these individuals may have been performing a different judgement task than the group as a whole. After applying this four-step data exclusion process, the remaining analyses were conducted on data obtained from 78 participants ($M_{Age} = 41.30$, $SD_{Age} = 11.32$, 45 Male, 32 Female, 1 No Response). Each robot was rated by between 12 and 36 participants.

B. Stimuli

Our stimuli set, depicted in Fig. 2, consisted of 251 still images of robots available in the ABOT Database, www.abotdatabase.info [11]. The ABOT Database is an online resource consisting of images of and data about real-world human-like robots that vary in the presence (i.e., number and salience) of human-like features. For every robot listed in the database, users can acquire an image of the robot depicted against a white or transparent background, in a standing, neutral, forward facing pose with a neutral or mildly positive facial expression (whenever possible). Accompanying each robot is an empirically derived overall human-likeness score. To detect confused or careless participants, we also included a stimuli set of "Catch trials" which consisted of 16 images: Eight images of humans that varied in age, gender, and ethnicity and eight images of featureless smart home devices (e.g., Amazon Echo, Google Home, Apple HomePod). The 16 images of humans and devices were the same as those employed in Phillips et al. [11].

C. Measure of human-likeness

For the current study, we borrowed from the ABOT data base an empirically derived overall physical human-likeness score that ranges from 0: *Not human-like* at all, to 100: *Just like a human*. We used these human-likeness scores for the 251 robots to quantify the "human-likeness" dimension of

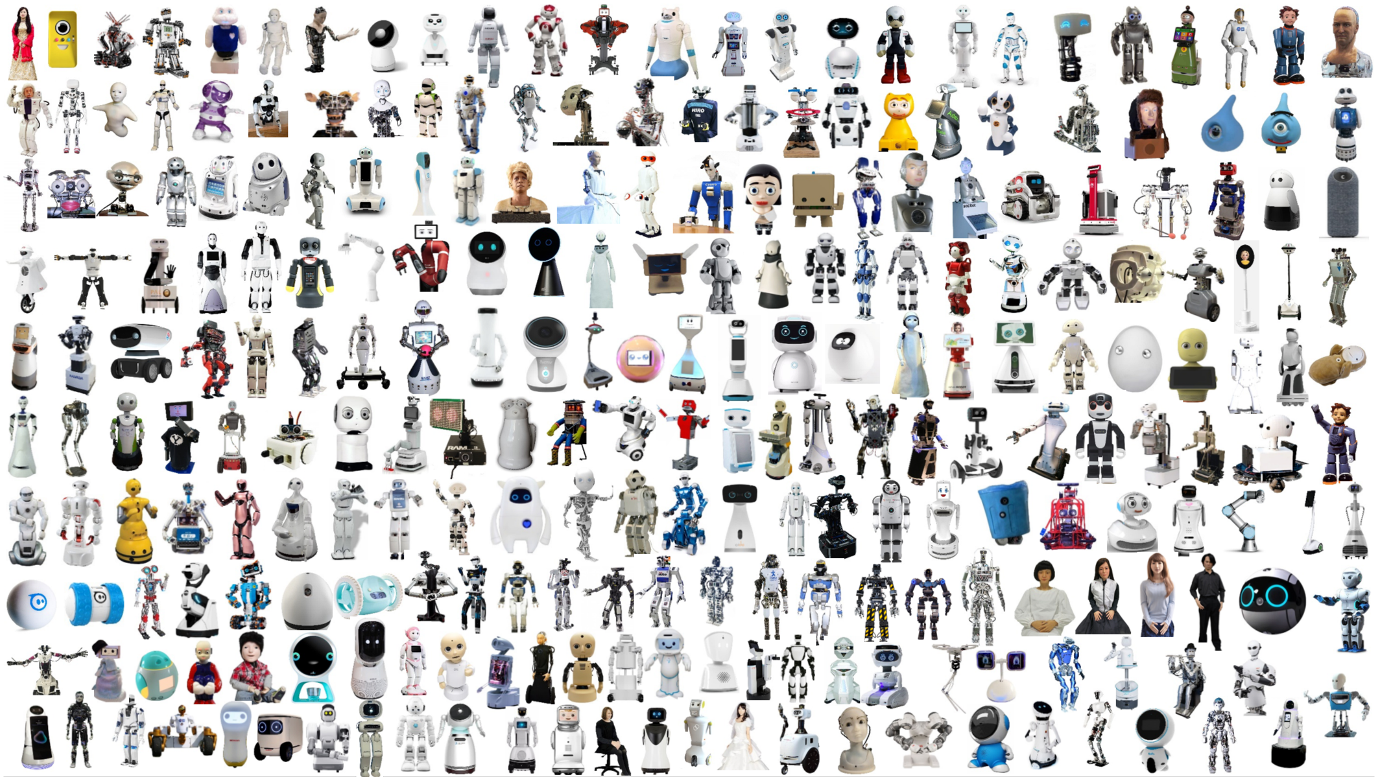

Fig. 2. All 251 human-like robots currently represented in the Anthropomorphic RoBOT (ABOT) Database

the uncanny valley hypothesis and to determine where in the human-likeness spectrum each robot in the database resides.

D. Measure of uncanniness

To encourage participants to use a consistent criterion in rating the stimuli on uncanniness, we provided them with a definition of uncanniness. This definition was derived from definitions found in the Oxford English and Merriam Webster's Dictionaries, and Dictionary.com. The following definition was provided to participants, uncanniness is: "The characteristic of seeming mysterious, weird, uncomfortably strange or unfamiliar." We then asked participants to rate each entity on how uncanny they perceived it to be using a slider ranging from 0 (*Not at all uncanny*) to 100 (*Extremely uncanny*).

E. Design and procedure

Two hundred and fifty-one images of robots were randomly assigned into one of four blocks of 50 images, or a fifth block of 51 robot images. All 16 images used for the catch trials were also added to each of the 5 blocks of images. Each participant was randomly assigned to judge all the images in one of these 5 blocks. Thus, each participant judged either 66 or 67 images of robots, people, and devices. For each trial in their block, one of the 50 or 51 robot images or one of the 16 images of humans and devices, appeared in the middle of the screen. Below the robot image appeared the probe question, "How uncanny is this?" Participants then used a slider to indicate their judgement for each of the robots in the image. For catch trials, the participants were asked to set the slider within a designated range on the scale (i.e., "Please drag the slider between x and y"). The 16 catch-trial images of humans and featureless devices were randomly intermixed with the images of robots and each image in the block was presented to participants at random. After judging all the images, participants were asked to complete a demographics questionnaire which asked participants to report their age, gender, native language, level of completed education, and prior knowledge of the robotics domain and experience working with robots. The entire study took between approximately $5 - 7$ minutes to complete. Participants were compensated $1 in return for their participation. The study protocol was approved by the Institutional Review Board of George Mason University.

III. Results

A visual inspection of the participants' responses to changes in the robots' human-likeness revealed the potential existence of not only one but two uncanny valleys (Fig. 3). Consistent with Mori's [1, 2] uncanny valley hypothesis, one large valley was present when the robots closely resembled humans, (i.e., human-likeness scores between 70 and 90). Most interestingly, another, smaller, valley was observed when robots had moderately little physical resemblance with humans (i.e., human-likeness scores between 10 and 30). These findings suggest that, even when the robots have a low or moderate resemblance with humans, if there are perceptual mismatches between different appearance dimensions in the robots, people may perceive them as uncanny.

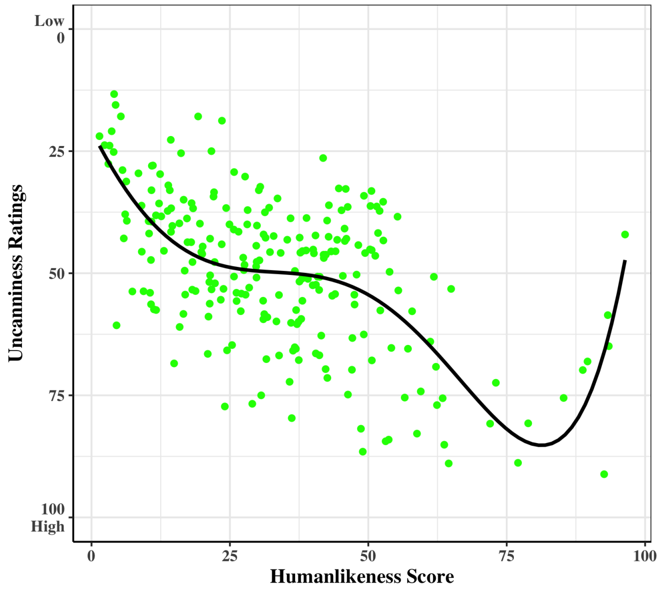

Fig. 3. Scatterplot depicting two valleys. The Y-axis of the graph representing uncanniness scores, has been reversed to depict the characteristic valleys as represented by high uncanniness.

To further verify the presence of either one or two valleys, we fit a polynomial mixed effects model on uncanniness scores by including the ABOT database's human-likeness scores as a fixed effect. We also included each participant as a random effect in order to control for participant-level variability in the model. We performed a model comparison analysis to examine which of a 3rd, 4th, and 5th order polynomial mixed effects model had the best fit to the dataset. Using the likelihood ratio tests, we assessed the goodness of fit of the models. Among these three possible polynomial mixed effects models, we predicted that the presence of Mori's [1, 2] uncanny valley would be determined by the 3rd order polynomial mixed effects model [15]. We also predicted that, if there was at least more than one valley, either the 4th or the 5th order polynomial mixed effects models would best explain the variances in the participants' ratings. Specifically, if the 4th order or the 5th order model had the best fit to the data, it would indicate the existence of the two uncanny valleys.

The model comparison results confirmed the 5th order polynomial mixed effects model as the best-fitting model, $\chi^2(1) = 118.02$, $p < .001$ for the 3rd vs. 4th order model; $\chi^2(1) = 8.28$, $p = .004$ for the 4th vs. 5th order model. These results indicate the presence of, not one, but two uncanny valleys. In the model, the effect of the 3rd order term of human-likeness was not significant ($p = .17$), but the effects of the 4th order term ($b = 263.33$, $\beta = .14$, $t = 10.94$, $p < .001$) and the 5th order term of human-likeness were significant ($b = 69.57$, $\beta = .04$, $t = 2.88$, $p = .004$).

IV. Discussion

With the largest set of real-world human-like robots to date, we found evidence supporting Mori's [1] uncanny valley. This valley was noticeable in participants' perceived uncanniness of 251 robots' which varied widely in the range and constituent features of human-likeness. Surprisingly, we found evidence of another, second uncanny valley for robots that had a moderately weak resemblance with humans. This discovery of the second uncanny valley was not previously predicted by the original uncanny valley hypothesis, nor found in the existing literature to date.

The present findings pose questions about the user experience design processes involved in creating the existing robots. We discovered that the novel uncanny valley emerged when the robots had human-likeness scores ranging between approximately 10 and 30 on the scale of 0 to 100. Out of 251 robots, 94 robots (37%) fell into this narrow range of robots. A common recommendation in the design community is to purposefully keep the human-like appearance of robots low, so as to diminish uncanny feelings towards, and false assumptions about robots intended to be social partners in homes and other human spaces [16]. However, it appears that the robot design community has been only moderately successful in creating robots with low human-likeness that are not uncanny. The emergence of the second valley implies that their recommendation to keep the human-like appearance of robots low will likely need to be more nuanced than originally conceived.

In contrast, there were only seven robots (0.3%) residing within the range of human-likeness scores between 70 and 90, where the original uncanny valley is hypothesized. It is likely that making truly human-like robots above 70 on the spectrum of human-likeness is difficult, explaining why there are so few robots in this higher range. With technical advances in social robot design and materials, future social robots may become much more human-like and hopefully have enough human-likeness to *jump* the valley.

Whether our observation that there are a nontrivial number of robots inducing the new uncanny valley is an outcome of robot designers' intentional decision processes, or a coincidental outcome remains to be answered. As the finding of the second uncanny valley is very novel, future work on replicating the findings with a larger sample size and investigating which psychological and design factors drive this second valley will be essential.

V. Conclusion

The human-like design of robots has been linked to a number of psychological constructs and mechanisms including trust [17, 18] empathy [19], collaboration [20], perceptions [21], compliance [22], and social interactions [23], among others. As robots are becoming more and more useful in today's world, a lot of time and interest is being invested into the appearance of these robots. A large emphasis has been put on making robots appear more [or less] human-like; thus increasing their likability and acceptance in human spaces.

Although the uncanny valley has been tested by a number of researchers, typically the exemplar robots used in these studies have not been comprehensive enough to find strong support for the proposed valley. Our work presented here to uncover the uncanny valley more precisely is important because it can help us gain a better understanding of the specific robots and their specific human-like configurations that lead to perceptions of uncanniness. A detailed understanding of these mechanisms will help designers to better match robot system design with intended purpose and context.

VI. Acknowledgment

This work is supported by Air Force Office of Scientific Research Grant 16RT0881f. We would also like to acknowledge the ABOT Database for making their stimuli freely available at www.abotdatabase.info, with documentation found in Phillips, E., Zhao, X., Ullman, D., Malle, B. F. (2018, February). What is Human-like? Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 105-113).

References

[1] Masahiro Mori. "The uncanny valley". In: *Energy* 7.4 (1970), pp. 33–35.

[2] Masahiro Mori, Karl F MacDorman, and Norri Kageki. "The uncanny valley [from the field]". In: *IEEE Robotics & Automation Magazine* 19.2 (2012), pp. 98–100.

[3] Kyoshiro Sasaki, Keiko Ihaya, and Yuki Yamada. "Avoidance of Novelty Contributes to the Uncanny Valley". In: *Frontiers in Psychology* 8 (2017). Publisher: Frontiers. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01792. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01792/full (visited on 04/02/2020).

[4] Carl F DiSalvo et al. "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. 2002, pp. 321–326.

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

[6] Karl F MacDorman and Hiroshi Ishiguro. "The uncanny advantage of using androids in cognitive and social science research". In: *Interaction Studies* 7.3 (2006), pp. 297–337.

[7] Christoph Bartneck et al. "Is the uncanny valley an uncanny cliff?" In: *RO-MAN 2007-The 16th IEEE international symposium on robot and human interactive communication*. IEEE. 2007, pp. 368–373.

[8] Chin-Chang Ho and Karl F MacDorman. "Measuring the uncanny valley effect". In: *International Journal of Social Robotics* 9.1 (2017), pp. 129–139.

[9] Megan Strait, Heather L Urry, and Paul Muentener. "Children's responding to humanlike agents reflects an uncanny valley". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 506–515.

[10] Astrid M Rosenthal-Von Der Pütten and Nicole C Krämer. "How design characteristics of robots determine evaluation and uncanny valley related responses". In: *Computers in Human Behavior* 36 (2014), pp. 422–439.

[11] Elizabeth Phillips et al. "What is human-like?: Decomposing robots' human-like appearance using the anthropomorphic robot (abot) database". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 105–113.

[12] Jari Kätsyri et al. "A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness". In: *Frontiers in psychology* 6 (2015), p. 390.

[13] Karoline Mortensen and Taylor L Hughes. "Comparing Amazon's Mechanical Turk platform to conventional data collection methods in the health and medical research literature". In: *Journal of General Internal Medicine* 33.4 (2018), pp. 533–538.

[14] Douglas J Ahler, Carolyn E Roush, and Gaurav Sood. "The micro-task market for lemons: Data quality on Amazon's Mechanical Turk". In: *Meeting of the Midwest Political Science Association*. 2019.

[15] James C Thompson, J Gregory Trafton, and Patrick McKnight. "The perception of humanness from the movements of synthetic agents". In: *Perception* 40.6 (2011), pp. 695–704.

[16] Evan Ackerman. "Ces 2017: Why every social robot at ces looks alike". In: *IEEE spectrum* (2017).

[17] Ewart J De Visser et al. "Almost human: Anthropomorphism increases trust resilience in cognitive agents." In: *Journal of Experimental Psychology: Applied* 22.3 (2016), p. 331.

[18] Peter A Hancock et al. "A meta-analysis of factors affecting trust in human-robot interaction". In: *Human factors* 53.5 (2011), pp. 517–527.

[19] Laurel D Riek et al. "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. 2009, pp. 1–6.

[20] Jennifer Goetz, Sara Kiesler, and Aaron Powers. "Matching robot appearance and behavior to tasks to improve human-robot cooperation". In: *Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003. The 12th IEEE International Workshop on*. Ieee. 2003, pp. 55–60.

[21] Kerstin S Haring et al. "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. 2016, pp. 265–270.

[22] Kerstin S Haring et al. “Robot authority in human-machine teams: effects of human-like appearance on compliance”. In: *International Conference on Human-Computer Interaction*. Springer. 2019, pp. 63–78.

[23] Brian R Duffy. “Anthropomorphism and the social robot”. In: *Robotics and Autonomous Systems* 42.3 (2003), pp. 177–190.