# Shape of the Uncanny Valley and Emotional Attitudes Toward Robots Assessed by an Analysis of YouTube Comments

Dawid Ratajczyk[1]

## Abstract

The uncanny valley hypothesis (UVH) suggests that almost, but not fully, humanlike artificial characters elicit a feeling of eeriness or discomfort in observers. This study used Natural Language Processing of YouTube comments to provide ecologically-valid, non-laboratory results about people's emotional reactions toward robots. It contains analyses of 224,544 comments from 1515 videos showing robots from a wide humanlikeness spectrum. The humanlikeness scores were acquired from the Anthropomorphic roBOT database. The analysis showed that people use words related to eeriness to describe very humanlike robots. Humanlikeness was linearly related to both general sentiment and perceptions of eeriness——more humanlike robots elicit more negative emotions. One of the subscales of humanlikeness, Facial Features, showed a UVH-like relationship with both sentiment and eeriness. The exploratory analysis demonstrated that the most suitable words for measuring the self-reported uncanny valley effect are: 'scary' and 'creepy'. In contrast to theoretical expectations, the results showed that humanlikeness was not related to either pleasantness or attractiveness. Finally, it was also found that the size of robots influences sentiment toward the robots. According to the analysis, the reason behind this is the perception of smaller robots as more playable (as toys), although the prediction that bigger robots would be perceived as more threatening was not supported.

**Keywords** Uncanny valley · Robot · YouTube · Emotions · Text mining · Social robotics · Natural Language Processing

## 1 Introduction

The uncanny valley hypothesis (UVH) suggests that almost, but not fully, humanlike artificial characters will elicit a feeling of eeriness or discomfort in observers (see [35,47]). This characteristic drop in likability is called the uncanny valley. Such an effect is considered within the field of humanoid robotics and also for computer-generated imagery (e.g., [20]). The concept of the uncanny valley has gained much attention in recent years; however, there are still certain inconsistencies in the debate. The doubts address not only the explanations for and the depth of the uncanny valley but also the dependent variable, which is commonly (but not exclusively) referred to as the affinity dimension [25]. These issues are strongly related to ambiguities with the emo-

tions related to the uncanny valley [25,62]. Kätsyri et al. [25] suggest that terms used in the uncanny valley studies (eeriness, likability, familiarity, and affinity) are related to various aspects of perceptual familiarity and emotional valence, and that the "empirical studies would be necessary for resolving which self-report items would be ideal for measuring affinity" (p. 3). There have been a few attempts to disambiguate self-report language describing the emotions in the uncanny valley within experimental studies (e.g., [20,21]); however, the results are not conclusive, and the uncanny valley research constantly does not use consistent language. The dependent variable has variously been considered as, for example, perceived warmth [27], eeriness, pleasantness, and creepiness [26], acceptance [54], or comfort level [30,43]. For a wider discussion, see Wang et al. [62, pp. 398–399].

Another point is that explanations of the uncanny valley focus robustly on the visual aspects of robots (e.g., [7,8,31,41,47]). However, recent research shows that negative or positive emotions toward identical artificial agents can be moderated by the individual's belief as to whether the agents are directed by artificial or human intelligence [50].

✉ Dawid Ratajczyk
dawid.ratajczyk@amu.edu.pl

1  Faculty of Psychology and Cognitive Science, Adam Mickiewicz University in Poznan, Szamarzewskiego 89/AB, 60–568 Poznań, Poland

Given the lack of agreement on what causes the uncanny effect, the involvement of variables additional to visual aspects (such as movement, behavioral social cues, proximity of agent, and others) seems plausible.

Most studies regarding UVH try to elicit the emotions in experimental conditions and measure them using ad hoc questionnaires (e.g., [39]) or unspecific, general-feeling questionnaires (e.g., [10]). Such an approach, despite the obvious advantages of controlled experiments, has negative implications for studying emotions. Aside from the explicit influence of an experiment on participants' emotions, the mood in which subjects walk into the laboratory has a large effect on positive and negative affect induction [14] and this may potentially lead to distorted results.

The sinusoidal UVH relationship introduced by Mori [35] is not unambiguously reflected in empirical research. The name of the UVH phenomenon (i.e., the "valley") refers to the specific graph shape of the emotional reaction toward humanlike agents – sharp decrease and increase of affinity. However the empirical evidence for such a curve is sparse. An extensive review by Kätsyri et al. [25] demonstrated a linear relationship between affinity and the humanlikeness, i.e., affinity reaction increased proportionally to the humanlikeness. A few other graphs have been considered, i.e., U-shaped relationship (e.g., [29,45]), and cliff-like relationship (e.g., [3]). Since the UVH was defined by the shape of the relationship between humanlikeness and affinity, there is a strong need to resolve this issue.

Considering the above, the reasons for lack of agreement in UVH research may be divided into: (a) inconsistencies in affinity dimension assessment, (b) limitation of research stimuli to visual aspects; and (c) difficulties in emotion elicitation in laboratory conditions. To address these issues I tested emotions and language in more ecological, non-laboratory conditions using Natural Language Processing (NLP) of comments on robots videos in social media. The novelty of the approach relies on studying more natural human reactions than those in surveys and experiments, as well as objectivization of the emotion assessment method. An analysis of large samples of natural human expressions using automated text processing allows to evaluate how various dependent variables (i.e., eeriness, pleasantness, and attractiveness) are associated with humanlikeness, possibly resulting in unraveling of some inconsistencies (addressing issue (a)). The present study exploits various videos of robots, hence it takes into account variables that are absent in 2D stimuli presentation, such as behavior of robots, their voice, size and others (addressing issue (b)). In contrast to experimental conditions, people using the Internet as part of their everyday life may manifest more natural emotional reactions toward the robots that they encounter. Therefore, an analysis of such manifestations allows to research genuine UVH-related reactions (addressing issue (c)).

During the last few years, social media has grown to become highly popular, providing not only a means of communication between people but also seizing control of many more social activities, such as the creation of reputation or the enabling a social life [36]. In order to study attitudes toward robots in popular media, I used comments on robot videos from the YouTube video-sharing platform. YouTube is a highly popular internet service (ranked #2 in global internet engagement according to *alexa.com*[1] ). Previous studies show that the analysis of YouTube comments allows community opinions and also emotions toward a specific topic to be determined [18,49]. One of the methods used to study affective states in people's statements is sentiment analysis. The method allows information to be gathered about attitudes, emotions, and opinions, and it is widely used in social media data extraction [1,5].

The structure of this paper is as follows. In the further part of the Introduction, I present studies that are relevant to the research objectives and formulate research hypotheses. In the Method section, I describe acquisition and processing of comments, and also selection of emotional and humanlikeness indicators. In the Results section, I test the hypotheses and perform additional exploratory analysis. In the Discussion, I interpret the results, generalize the findings, and consider the limitations of the study. The paper is supplemented with an Appendix, where the list of all analyzed robots and their scores is presented.

## 1.1 Related Work

Among the work with the closest methodological approach to this research, several studies have examined online comments related to the evaluation of robots. One of the first ones in this field was carried out by Friedman et al. [15]. They investigated online discussion forums associated with robotic dog AIBO. During the study, 3119 postings were coded with subsequent general categories: technological essences, life-like essences, mental states, social rapport, and moral standing. Results showed that AIBO psychologically engaged its owners, and people created relationships with the robotic pet. The robot evoked conceptions of life-like essences, mental states, and social rapport, and sporadically conceptions of moral standing. The authors considered that playfulness might have a part in language people use to refer to their robotic pets, as it influences users actual emotions and thoughts.

Strait et al. [52] performed an analysis of YouTube comments regarding robots using the raters method, which involved an evaluation of comments' topics by assistants. They investigated public perception toward groups of

---

[1] https://www.alexa.com/siteinfo/youtube.com                    [Accessed: 07.12.2020].

mechanomorphic and humanlike robots and established a less positive valence and more UVH-related comments toward humanlike robots compared with mechanomorphic ones. They also discovered that decreased emotional valence in comments is partially related to fears of a "technology takeover" (i.e., a fear of robots replacing humans). Additionally, they stressed the occurrence of the widespread sexualization and objectification of the female-gendered robots in comments. However, this study has certain limitations. Their analysis involved comments from subjectively chosen videos and also the subjective exclusion of some comments. The number of comments analyzed by judges was rather small (1200) and their methodology did not enable them to perform a systematic inquiry of emotional words used for the description of uncanny characters.

Hover et al. [22] extended the work of Strait et al. [52] and used a similar methodological approach to the analysis of comments regarding more and less humanlike robots. Their results also indicated that more humanlike robots elicit more negative emotional reactions related to the uncanny valley effect. They also confirmed the results that female humanlike robots were more likely to be subject to sexualization and sexism than male robots. Their data suggests that there is more difference in examined factors between less and more humanlike robots than female and male robots. Also, less humanlike robots were more likely to evoke perceptions of a threat than highly humanlike robots.

Strait et al. [51] tested racialization toward robots with appearances of different racial identities based on analyses of YouTube comments. They used comments from videos of a Black, White, and Asian appearing robots and humans (6 videos in total). The results show that people extend and amplify racial biases toward robots, and also that dehumanization based on social stereotypes is greater for robots.

Vlachos and Tan [59], instead of using human annotators, did an analysis of YouTube comments regarding four humanlike robots (involving comments on four videos), with the utilization of text mining and machine learning. Their work was entirely exploratory, not focused explicitly on the uncanny valley, but rather on a general interaction with highly humanlike robots. The authors distinguished three topics important for robotics: human–robot relationships, technical specifications, and the so-called science fiction valley (a combination of the UVH concept and references to science fiction movies and games). The limitations of this study were the choice of only four videos, a lack of manipulation of humanlikeness, and the involvement of replies to main comments, which may contain off-topics.

Also, Yu [63] studied attitudes toward robots employed as hotel workers. They collected comments on two YouTube videos and coded them automatically in reference to concepts related to the perception of robots (anthropomorphism, animacy, likability, perceived intelligence, and perceived safety). Their cluster analyses of data showed that likability and anthropomorphism are the most distinct concepts. The results supported the existence of the uncanny valley. The discomfort in form of anxiety feeling co-occurred with perceived intelligence. Additionally, discussions about movement of robots were related to machinelikeness. Although the thematic analysis was automatic, the cleaning of data was done manually. Therefore, the sample size of comments and videos was rather small and videos were limited to robots from a very specific context.

Considering the variables, that may have influence on robot perception, the results of Wang et al. [61] show that the size of agents which are otherwise visually equivalent determines the degree to which they are perceived as uncanny. Participants in augmented reality preferred smaller virtual agents over visually identical human-size agents, referring to these as too large, imposing, weird, and creepy. These findings are in line with the conclusions of the analysis of Kätsyri et al. [25] pointing out that the uncanny valley concept is in fact very complex and suggest there is a need for a closer examination of the influence of robot size on UVH-related feelings. Also, Mori [35] pointed that higher humanlikeness may be perceived when absolute size of agent is ignored. Wang et al. [61] suggested, on the bases of subjects' feedback, that small embodied agents are more entertaining and amusing than other agents. Also, Wagner et al. [60] reported that fun plays an important role in embodied agents. As such, people may treat smaller robots as if they were playable or related to fun. Another plausible explanation is that bigger robots can be seen as stronger and more threatening to people, and therefore evoke negative emotions toward them.

In the following study, I use methods of data acquisition and processing that are automated and resilient to the possibility of a biased, subjective choice of movies and comments. I acquire a large number of utterances referring to robots with different humanlikeness and conduct several analyses in order to exploit the potential of the data.

The aims of this paper are as follows: (1) to test the relationship between robot humanlikeness and sentiment scores, (2) to examine which of the variables (eeriness, pleasantness, and attractiveness) are related to humanlikeness with the new NLP method, (3) to test the impact of robot size on sentiment and to examine the reasons behind the observed relationship, and (4) to characterize the specific, emotional words expressed toward robots.

Additionally, I investigated the awareness of UVH among commenters. The popularity of the UVH concept among internet communities seems to be widespread, as indicated in popular articles and by an extensive list of spotted uncanny

valley examples in animations and video games prepared by users[2].

## 1.2 Hypotheses

On the basis of the above-mentioned literature, the following hypotheses have been formulated:

**H 1** *The shape of the graph representing the relationship between humanlikeness (and its subscales) and sentiment valence toward robots is linear.*

**H 2** *Emotional indicators (eeriness, pleasantness, and attractiveness) are equally related to humanlikeness. The relationships between humanlikeness (and its subscales) and emotional indicators are linear.*

**H 3** *The size (i.e., height) of robots has an impact on emotions elicited by robots.*

**H 3a** *The smaller a robot, the more it is perceived as playable or related to fun.*

**H 3b** *The bigger a robot, the more it is perceived as threatening and dangerous.*

## 2 Methods

The work included: (1) data retrieval (downloading the comments regarding robots from the YouTube platform), (2) processing of comments (cleaning the text and extracting emotional indicators), and (3) acquiring humanlikeness scores for robots.

### 2.1 YouTube Comments Collection

The method of data collection was inspired by the publication of Thelwall [55]. The method allows videos relevant to a given topic to be systematically searched for and acquired without engaging subjective preferences.

The topic of the investigation was existing robots. In order to acquire utterances regarding robots from a humanlikeness spectrum, I prepared a list of 246 developed and functional robots—242 robots from the Anthropomorphic roBOT (ABOT) Database[3] plus 4 additional robots which were not included in the ABOT database[4].

I adopted the API protocol shared by YouTube and wrote a Python 3.8 script in order to download comments related to a particular robot. Firstly, I acquired the relevant videos list for each robot. The criteria of the search for videos were as follows. I included only short videos (less than 4 min[5]) in order to focus on robots' presentations and reduce the possibility of the occurrence of uncontrolled variables such as the presentation of multiple robots or excessive commentary in the video, for example. The relevance language (the API option) was English and the region of the search was the US[6]. Videos (and comments) were not limited by date. The search phrase was combined from the robot's name, the word 'robot' and the additional clue (the name of the production company or creator, country where the robot was developed, or the word 'humanoid'). The phrases were prepared in order to maximize the number of relevant videos. All the search phrases are included in the Supplementary Materials[7]. The comments scraping procedure was performed between 1st and 10th August 2020.

After selecting videos with the described method, I automatically evaluated their accuracy according to the following criterion: I left only the videos (2157 in total from original 8782) which had the name of the given robot in the title[8]. After this, I downloaded all of the comments for the listed videos. Concerning the relevance to the topic of the videos (robots), I discarded replies to comments and left only primary comments. I removed empty comments and those duplicated with more than 100 characters (long duplicated comments might be spam or created by bots). Then I removed non-English comments, which were detected with the use of the Python *langdetect* v1.0.7 package (https://pypi.org/project/langdetect/). The total number of comments after processing was 228,688 from 2149 videos. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

### 2.2 Analysis of Comments

Each comment was processed with Python 3.8 script[9]. Firstly, all of the hyperlinks were removed from the comments and the comments were part-of-speech tagged with the use of the *NLTK POS tagger*. Then the stopwords, punctuation,

---

[2] https://www.wired.com/story/uncanny-valley-robot-voices/; https://tvtropes.org/pmwiki/pmwiki.php/Main/UncannyValley [Accessed: 07.12.2020].

[3] http://abotdatabase.info/collection [Accessed: 31.08.2020]; 9 robots were excluded from the original 251 item list due to nonspecific names and potential collisions during YouTube search.

[4] Robots added: Vector by Anki, Spot by Boston Dynamics, earlier and later version (after 28.03.2019) of Handle by Boston Dynamics. They were added in order to maximize the number of robots for analysis.

[5] The YouTube API allows videos below 4, between 4 and 20, and above 20 minutes to be searched for.

[6] Due to the YouTube API features, this does not exclude all the other videos (with different languages or from different regions).

[7] The Supplementary Materials are available at https://osf.io/cyvmx/.

[8] The reason behind this is that the YouTube search algorithm is not straightforward and may decide that a video is relevant even if there is no keyword in the title.

[9] The following packages were used: *string*, *NLTK*, *AFINN*, *collections*, *NumPy*, and *pandas*.

**Table 1** Example of counting the sentiment scores for comments with *AFINN* package

| | Original comment | Processed comment with word scores | Score |
|---|---|---|---|
| 1. | "This is awesome. Amazing!" | Awesome [4], amazing [4] | 8 |
| 2. | "It'll be scary when they put guns in their hands" | Scary [− 2], put [0], gun [− 1], hand[0] | − 3 |
| 3. | "It's ridiculous, rough, ugly and so more adjectives." | Ridiculous [− 3], rough [0], ugly [− 3], adjective [0] | − 6 |
| 4. | "Does that robot vaccum too?" | Robot [0], vaccum [0] | 0 |

and non-alphabetic words were removed. All the remaining words were lemmatized.

In order to obtain reliable data for further analysis, I used only those robots that have more than 200 suitable (not discarded by the above criteria) comments in total. The same cut-off number was used by Guzman et al. [17] in a similar sentiment analysis of comments from the GitHub platform. The selected number is a trade-off between an insufficient number of comments for unbiased analysis and a sufficient number of robots for further analysis (see elaboration in the Limitations section). That left 33 robots suitable for further analysis (224,544 comments from 1515 videos in total). List of these robots is presented in the "Appendix".

*Sentiment score* The sentiment scores were counted with the use of the *AFINN* v0.1 Python package (https://pypi.org/project/afinn/), which provides a lexicon of emotional words with scores ranging from − 5 to 5. The lexicon was prepared on the basis of internet fora and microblogs, and contains internet slang and obscene words [37], making it suitable for a YouTube comments analysis. The score for each comment was counted by adding up individual scores for every word in the processed comment (see example in Table 1). The mean from all comments referring to a robot was then counted in order to obtain the robot sentiment score. All the scores for individual robots are presented in the "Appendix".

*Eeriness, pleasantness, and attractiveness indices* Additionally, I have defined indices to distinguish emotional terms characteristic to the uncanny valley, i.e., related to emotions elicited by observation or contact with close to human-like agents. The reason for choosing concepts of eeriness, pleasantness, and attractiveness is that these are the most discussed dependencies in the context of UVH studies (e.g., [20,25,57]). There are tools available for accurate emotions identification in language such as Linguistic Inquiry and Word Count (LIWC) [53]. However, the area of the uncanny valley is very specific and, although there were attempts to disambiguate the words used for naming emotions toward robots at least for eeriness (e.g., [21,48]), in order to maintain consistency among examined concepts, I created original word sets for identifying UVH–related emotions. I created three lists of words related to each aforementioned concept

**Table 2** All the words counted for eeriness, pleasantness, attractiveness, and familiarity indices

| Index | Words related to the index |
|---|---|
| Eeriness | Creepy, eerie, haunting, spookish, spooky, uncanny, unearthly, weird |
| Pleasantness | Agreeable, blessed, congenial, darling, delectable, delicious, delightful, delightsome, dreamy, dulcet, enjoyable, felicitous, good, grateful, gratifying, heavenly, jolly, luscious, nice, palatable, pleasant, pleasing, pleasurable, pretty, satisfying, savory, sweet, tasty, welcome |
| Attractiveness | Aesthetic, attractive, beauteous, beautiful, bonny, comely, cute, drop-dead, fair, fetching, good, good-looking, goodly, gorgeous, handsome, knockout, likely, lovely, lovesome, pretty, ravishing, seemly, sightly, stunning, taking, well-favored |

using the Merriam-Webster Online Thesaurus[10]. As the list of videos acquired from YouTube was targeted for the US, I used the American dictionary, and this one was used previously by Kätsyri et al. [25] for defining the concepts related to the uncanny valley. The following definitions for each concept were used when identifying synonyms: eerie—fearfully and mysteriously strange or fantastic; pleasant—giving pleasure or contentment to the mind or senses; attractive—very pleasing to look at. Lists of all synonyms for each index are presented in Table 2. Afterward, eeriness, pleasantness, and attractiveness indices were counted for each robot. Establishing an index for a given robot required the following steps. I counted the relative frequencies of words related to each concept. I excluded the counts of the word 'uncanny' which occurred in the phrase 'uncanny valley' in order to focus on expressions of emotions, not the awareness of the phenomenon. This method gave me the possibility of a systematic, numerical evaluation of each concept for every robot. All the scores for individual robots are presented in the "Appendix".

---

[10] https://www.merriam-webster.com/thesaurus [Accessed: 07.12.2020].
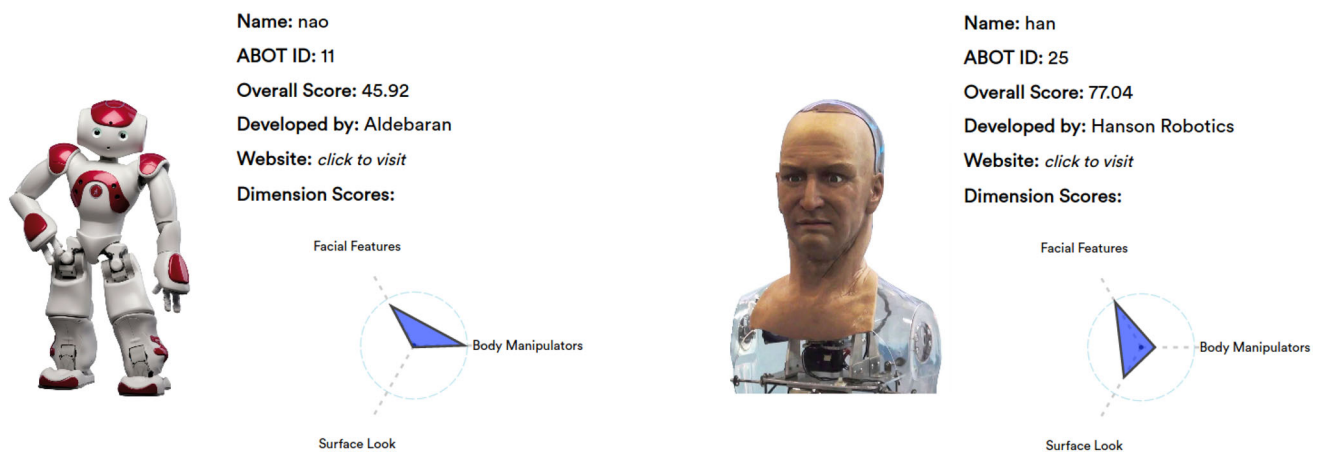
Fig. 1 Illustrative robots from ABOT database with humanlikeness (and subscales) scores

## 2.3 Humanlikeness of Robots

Appraisals of robots' humanlikeness were acquired from the ABOT database, and using a tool provided by the database creators [40]. The ABOT database is an open source collection of real-world robots with their humanlikeness scores. The database allows the unification of robots' humanlikeness dimension among studies and the investigation of the impact of distinguished underlying factors.

Phillips et al. [40] created the ABOT database with humanlikeness scores and related factors based on a study using a collection of 200 images of real-world robots. They uncovered three distinct appearance dimensions (i.e., bundles of features) that contribute to the anthropomorphism of robots. They distinguished the following subscales: (1) Surface Look (presence of eyelashes, head hair, skin, genderedness, nose, eyebrows, apparel), (2) Body-Manipulators (presence of hands, arms, torso, fingers, legs), and (3) Facial Features (presence of face, eyes, head, mouth). The three subscales positively correlate with humanlikeness. Exemplary robots from ABOT database are presented in Figure 1.

The ABOT authors shared scores of general humanlikeness and its subscales for 251 robots and also made available a tool for assessing the humanlikeness of robots not present in the database. Because the ABOT scales were created on a large sample of participants (over 1000), it seems to be an appropriate appliance for the unification of robots' humanlikeness dimension among studies. These scores will be used for further analyses. All the ABOT scores for robots used for this study are presented in the Appendix.

In what follows, I will use the main ABOT humanlikeness score but also all of its subscales.

## 3 Results

### 3.1 Humanlikeness and Sentiment Scores

The UVH describes a non-linear relationship between humanlikeness and emotional reaction. Firstly, I tested the relationship between humanlikeness of robots and general sentiment scores and examined how the relationship changes for particular ABOT subscales of humanlikeness (H1). Therefore, in order to characterize the shape of the relationship and test which model better describes the relationship (linear, quadratic or cubic), I performed polynomial curve fitting and tested the goodness of fit of each regression model using the Akaike Information Criterion (AIC; see [9]). AIC allows us to compare models of varying complexity, penalizing a higher number of parameters. For small sample sizes ($n/K < 40$, where $n$ is sample size and $K$ is number of parameters), as in this case, Burnham and Anderson [9, p. 66]) suggested the usage of the adjusted formula:

$$AIC_c = n * \ln\left(\frac{RSS}{n}\right) + 2 * K + \frac{2 * K * (K + 1)}{n - K - 1},$$

where $RSS$ is residual sum of squares. A model with the lowest $AIC_c$ is preferred. Also, $R^2$ was counted in order to examine the variance in the dependent variable that is predicted from the independent variable. The results are presented in Table 3. $P$ values were corrected with the Benjamini–Hochberg adjustment (see [6]) for multiple comparisons (12 tests).

The results show that, out of the three models, the linear one was the one which best fit the relationship between humanlikeness and sentiment. As for the ABOT subscales of humanlikeness, Surface Look also has a linear relationship with sentiment. However, the relationship between Facial Features and sentiment was best represented by the cubic model. Sentiment was moderate at very low humanlikeness,

**Table 3** Polynomial regression comparison between models of the humanlikeness (and subscales) and sentiment relationship

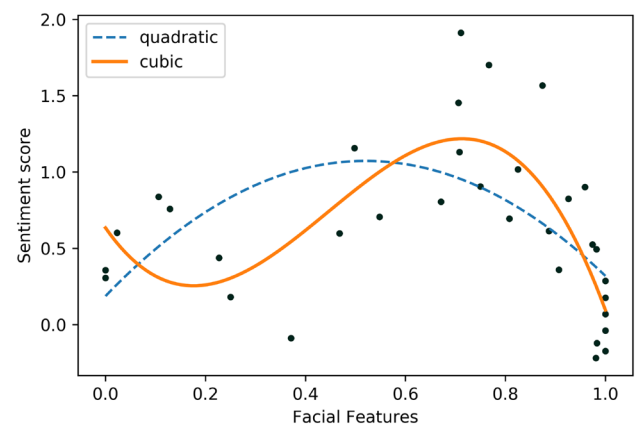| Dependent var. | Independent var. | Model | $AIC_c$ | $\Delta_i(AIC)$ | $R^2$ | $F$ | $p$ |
|---|---|---|---|---|---|---|---|
| Sentiment | Humanlikeness | **Linear** | **− 47.02** | **0.0** | **0.26** | **10.81** | **0.015** |
| | | Quadratic | − 45.05 | 2.0 | 0.27 | 5.52 | 0.022 |
| | | Cubic | − 43.38 | 3.6 | 0.29 | 3.93 | 0.027 |
| | Surface look | **Linear** | **− 46.2** | **0.0** | **0.24** | **9.78** | **0.015** |
| | | Quadratic | − 43.79 | 2.4 | 0.24 | 4.74 | 0.027 |
| | | Cubic | − 44.48 | 1.7 | 0.31 | 4.39 | 0.023 |
| | Facial features | Linear | − 37.25 | 22.3 | 0.0 | 0.1 | 0.757 |
| | | Quadratic | − 46.15 | 13.4 | 0.29 | 6.21 | 0.017 |
| | | **Cubic** | **− 59.52** | **0.0** | **0.56** | **12.51** | **<0.001** |
| | Body-Manipul. | Linear | − 38.01 | 0.0 | 0.03 | 0.82 | 0.497 |
| | | Quadratic | − 36.1 | 1.9 | 0.04 | 0.64 | 0.641 |
| | | Cubic | − 33.52 | 4.5 | 0.04 | 0.42 | 0.757 |

Bold, underlined rows indicate best, significant models for particular relationships
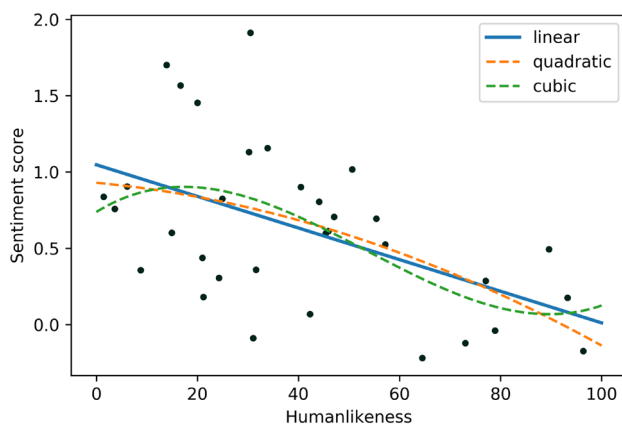


**Fig. 2** Sentiment score and Humanlikeness scale relationship curve fitting
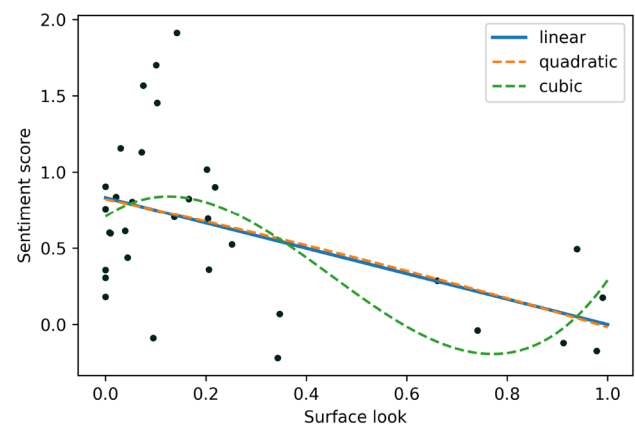


**Fig. 3** Sentiment score and Surface Look subscale relationship curve fitting

decreased at low humanlikeness, was highest at high humanlikeness, but lowest at very high humanlikeness. For the Body-Manipulators subscale, none of the models was significant. Plots for significant models are presented in Figs. 2, 3, and 4. The best models were drawn with a solid line. The results partially support H1—out of 4 examined relationships, 2 were linear. Generally, the more humanlike robots are (and the more humanlike surface they have), the more negative sentiment they elicit.

## 3.2 Humanlikeness and Emotional Indicators

In order to test the associations between particular emotional indicators (eeriness, pleasantness, and attractiveness) and humanlikeness (H2) I conducted a similar polynomial regression analysis as for the relationship between humanlikeness and sentiment. For each indicator the $AIC_c$ values, $R^2$, and significance values for models (with Benjamini–Hochberg



**Fig. 4** Sentiment score and Facial Features subscale relationship curve fitting

adjustment for 12 tests) were calculated. Results are shown in Table 4.

**Table 4** Polynomial regression comparison between models of the humanlikeness (and subscales) and emotional indicators relationships

| Dependent var. | Independent var. | Model | $AIC_c$ | $\Delta_i(AIC)$ | $R^2$ | F | p |
|---|---|---|---|---|---|---|---|
| Eeriness | | | | | | | |
| | Humanlikeness | **<u>Linear</u>** | **<u>66.12</u>** | **<u>0.0</u>** | **<u>0.28</u>** | **<u>11.94</u>** | **<u>0.003</u>** |
| | | Quadratic | 68.32 | 2.2 | 0.28 | 5.92 | 0.012 |
| | | Cubic | 68.78 | 2.66 | 0.33 | 4.72 | 0.013 |
| | Surface Look | Linear | 63.25 | 3.88 | 0.34 | 15.84 | 0.001 |
| | | Quadratic | 60.96 | 1.59 | 0.43 | 11.15 | 0.001 |
| | | **<u>Cubic</u>** | **<u>59.37</u>** | **<u>0.0</u>** | **<u>0.49</u>** | **<u>9.47</u>** | **<u>0.001</u>** |
| | Facial Features | Linear | 72.31 | 14.92 | 0.13 | 4.59 | 0.053 |
| | | Quadratic | 62.06 | 4.67 | 0.41 | 10.29 | 0.001 |
| | | **<u>Cubic</u>** | **<u>57.39</u>** | **<u>0.0</u>** | **<u>0.52</u>** | **<u>10.65</u>** | **<u>0.001</u>** |
| | Body-Manipul. | Linear | 76.76 | 0.0 | 0.0 | 0.11 | 0.839 |
| | | Quadratic | 78.72 | 1.97 | 0.02 | 0.26 | 0.839 |
| | | Cubic | 81.32 | 4.56 | 0.02 | 0.17 | 0.914 |
| Pleasantness | | | | | | | |
| | Humanlikeness | Linear | 35.76 | 0.04 | 0.14 | 5.13 | 0.146 |
| | | Quadratic | 35.72 | 0.0 | 0.2 | 3.84 | 0.146 |
| | | Cubic | 36.31 | 0.59 | 0.25 | 3.24 | 0.146 |
| | Surface Look | Linear | 37.86 | 0.0 | 0.09 | 2.91 | 0.295 |
| | | Quadratic | 39.11 | 1.25 | 0.12 | 2.0 | 0.298 |
| | | Cubic | 40.28 | 2.42 | 0.16 | 1.78 | 0.298 |
| | Facial Features | Linear | 40.62 | 0.0 | 0.01 | 0.18 | 0.671 |
| | | Quadratic | 41.1 | 0.48 | 0.06 | 1.01 | 0.412 |
| | | Cubic | 42.02 | 1.4 | 0.11 | 1.19 | 0.412 |
| | Body-Manipul. | Linear | 39.87 | 0.56 | 0.03 | 0.9 | 0.412 |
| | | Quadratic | 39.31 | 0.0 | 0.11 | 1.9 | 0.298 |
| | | Cubic | 41.91 | 2.6 | 0.11 | 1.23 | 0.412 |
| Attractiveness | | | | | | | |
| | Humanlikeness | Linear | 65.99 | 0.0 | 0.03 | 1.1 | 0.697 |
| | | Quadratic | 67.9 | 1.91 | 0.05 | 0.78 | 0.697 |
| | | Cubic | 69.19 | 3.2 | 0.09 | 0.91 | 0.697 |
| | Surface Look | Linear | 67.13 | 0.0 | 0.0 | 0.01 | 0.929 |
| | | Quadratic | 69.41 | 2.28 | 0.0 | 0.07 | 0.929 |
| | | Cubic | 69.38 | 2.25 | 0.08 | 0.85 | 0.697 |
| | Facial Features | Linear | 66.7 | 0.0 | 0.01 | 0.42 | 0.697 |
| | | Quadratic | 68.75 | 2.05 | 0.02 | 0.38 | 0.825 |
| | | Cubic | 69.54 | 2.84 | 0.08 | 0.8 | 0.697 |
| | Body-Manipul. | Linear | 63.37 | 0.0 | 0.11 | 3.76 | 0.697 |
| | | Quadratic | 65.78 | 2.41 | 0.11 | 1.83 | 0.697 |
| | | Cubic | 68.2 | 4.83 | 0.11 | 1.24 | 0.697 |

Bold underlined rows indicate best, significant models for particular relationships

The results indicate that for the general humanlikeness and eeriness relationship, the best model is linear. However, for the Surface Look and Facial Features subscales, the best models are cubic. As for pleasantness and attractiveness, none of the models was significant. Plots of significant models are presented in Figs. 5, 6, and 7. Therefore, the first part of H2 should be rejected—emotional indicators are not equally related to the humanlikeness. Eeriness is the most important factor for humanlikeness. The results partially support the second part of H2—the relationship between humanlikeness and the only significant index (eeriness) is linear. However, none of humanlikeness subscales affect the eeriness linearly.
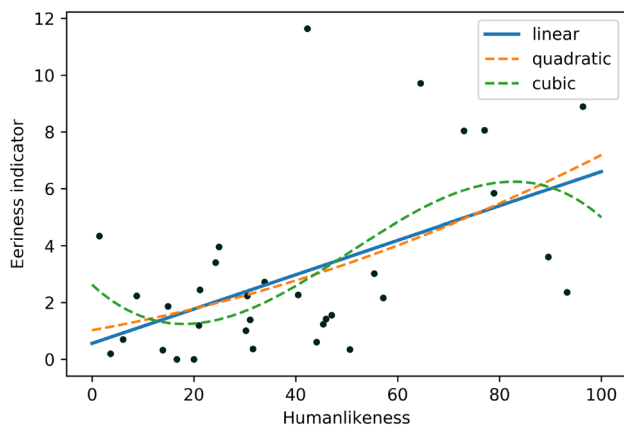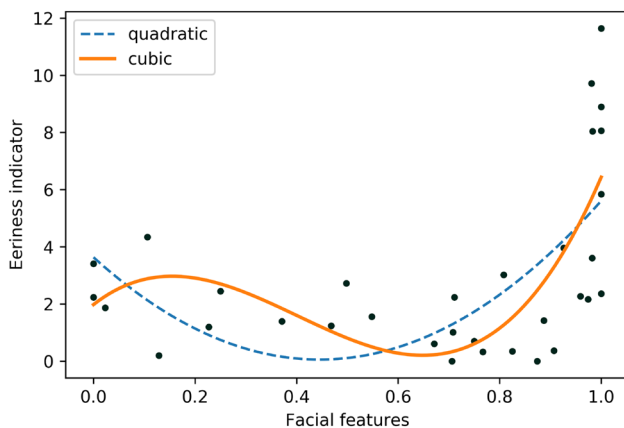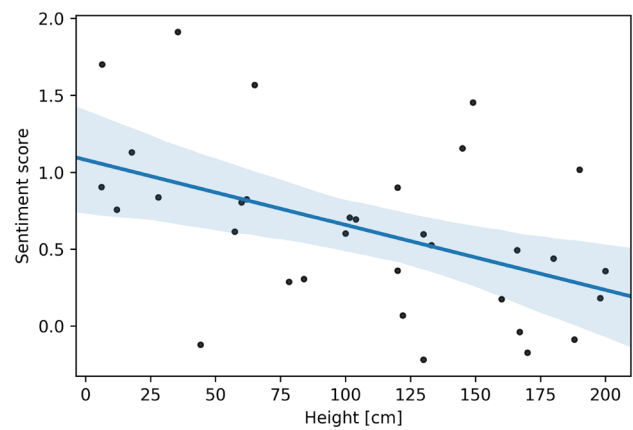
**Fig. 5** Eeriness index and Humanlikeness scale relationship curve fitting
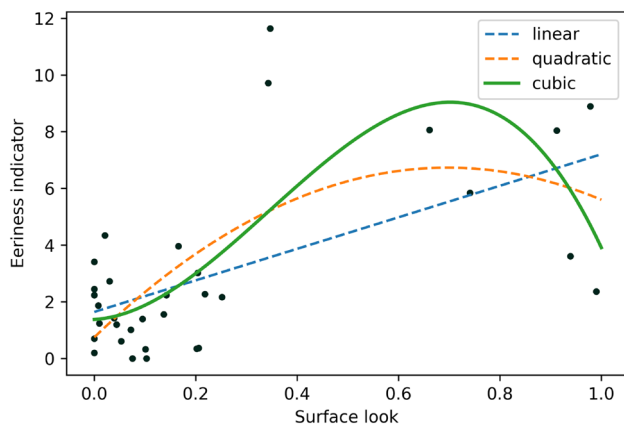


**Fig. 6** Eeriness index and Surface Look subscale relationship curve fitting



**Fig. 7** Eeriness index and Facial Features subscale relationship curve fitting



**Fig. 8** Regression plot of height of robots and sentiment score relationship

### 3.3 Height of the Robot and Sentiment Score

Literature regarding UVH subject suggests that the size of robots may have an impact on emotions elicitation [35,61]. It may be a confounding factor in the human–robot interaction (HRI) studies. Nevertheless, it seems that the relationship has not been tested empirically. Therefore, I tested the influence of robots' height on general sentiment score (H3). The size of robots was acquired from documentations shared by producers, promotional materials, and articles about robots. For two robots (Han, Bina48), the information about their size was not available, therefore I used multiple independent raters method to evaluate their height. 6 experts in social robotics scrutinized pictures and movies of the robots, and estimated the height on the basis of the comparison to various elements visible in the scenes (humans, computers, other elements that their found useful). The mean of their responses was taken as the height[11].

The regression model for predicting the sentiment by robot size was significant ($F(31, 1) = 9$, $p = 0.005$), $R^2$ was equal to 0.23. The size coefficient was equal to $\beta = -0.004$ ($t(31) = -3$, $p = 0.005$). The regression plot is shown in Figure 8. This result supports the H3, indicating that the height of a robot has a significant impact on general emotions elicited by robots. The bigger a robot is, the more negative emotions it elicits.

### 3.4 Explanation of Sentiment Score and Robots' Height Relation

In order to find out what is the reason behind the observed relationship between sentiment score and robots' height, hypotheses H3A and H3B have been formulated. H3A states

---

[11] Estimated height for Han: 78.3 cm ($SD = 14.6$); for Bina48: 44.2 cm ($SD = 4.5$).

**Table 5** All the words counted for the playfulness and threateningness indices

| Index | Words related to the index |
|---|---|
| Playfulness | Dalliance, frolic, frolicking, fun, play, recreation, relaxation, rollicking, sport |
| Threateningness | Dangerous, grave, grievous, hazardous, jeopardizing, menacing, parlous, perilous, risky, serious, threatening, unhealthy, unsafe, venturesome |



**Fig. 9** Regression coefficients for the relationship between height and sentiment score as mediated by playfulness. The regression coefficient between height and sentiment, controlling for playfulness, is in the parentheses. $*p < .05$, $***p < 0.001$

that people may treat smaller robots as if they were playable or related to fun (see [60,61]). Also, developed on the basis of the intuition that bigger robots can be seen as stronger and more threatening to people, H3B states that the decreased sentiment for bigger robots is related to perceived threateningness.

To test these explanations I defined two additional indices: 'playfulness index' and 'threateningness index', analogously to previous uncanny indices. I took the synonyms from the Merriam–Webster Online Thesaurus of the word 'play' in the following meaning: "activity engaged in to amuse oneself", and of the word 'threatening' in the meaning: "involving potential loss or injury". The synonyms used for the indices are presented in Table 5.

Subsequently, I then conducted mediation analysis, testing if playfulness and threateningness indices mediate the relationship between the height of robots and sentiment scores. The analysis was performed using R software [42] and the *mediation* package [56]. Using Baron and Kenny's [2] procedure, I tested the influence of height (independent variable) on the playfulness and threateningness indices (mediators) separately. The model for playfulness was significant ($F(31, 1) = 6.4$, $p = 0.017$), and the model for threateningness was not significant ($F(31, 1) = 2.9$, $p = 0.1$). Therefore, I did not find evidence to support the hypothesis that threateningness is a mediator of the height and sentiment relationship. Next, I tested the combined influence of height and playfulness on sentiment. The effect of size on sentiment became no longer significant ($t = -1.8$, $p = 0.09$), while the effect of playfulness remained significant ($t = 3.3$, $p = 0.003$). Mediation schema with coefficients is presented in Fig. 9. The bootstrapping test with 10,000 simulations showed that the mediation was significant ($ACME = -0.0018$, CI [$-0.0034$, 0.0], $p = 0.044$). Therefore, playfulness was found to be a significant mediator of the height and sentiment relationship. H3A is supported and smaller robots are perceived as more playable (as toys). H3B is rejected and it cannot be confirmed that people perceive bigger robots as more threatening.
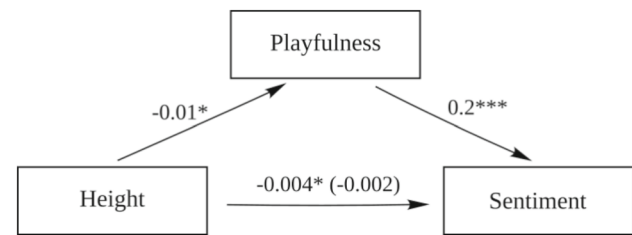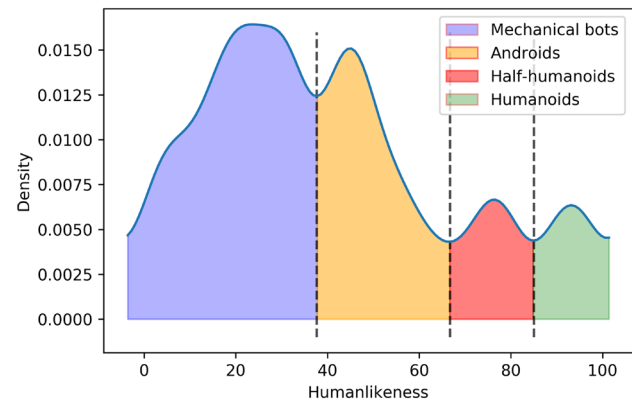


**Fig. 10** The kernel density estimate plot with distinguished groups

### 3.5 Exploratory Analysis

The following analysis is conducted in order to define the most suitable words for measuring self-reported attitudes toward robots. I grouped robots with similar humanlikeness scores (from ABOT database) to examine how people describe them. I used the kernel density estimation in order to distinguish groups within humanlikeness scale (gaussian kernel, bandwidth equal to 5)[12]. Four groups were distinguished: (1) mechanical bots – low humanlike robots with mechanistic surface, and low to medium humanlike facial features and body [humanlikeness score: 0–37]; (2) androids— medium humanlike robots with facial and bodily features, but with low humanlike surface [humanlikeness score: 37–67]; (3) half-humanoids—humanlike robots with surface, and facial features resembling humans, but without entirely humanlike body [humanlikeness score: 67–85]; and (4) humanoids— highly humanlike robots with humanlike surface, facial and bodily features [humanlikeness score: 85–100]. The kernel density estimate is presented in Fig. 10. Table 6 shows means

---

[12] Usage of the optimal bandwidth calculated according to Silverman's rule of thumb ($bw = 11.8$) resulted in one data cluster, therefore I had been decreasing the bandwidth until 4 groups were distinguished, as four is the minimal number of humanlikeness groups to describe the uncanny valley [25].

**Table 6** Average humanlikeness and its subscales scores for distinguished groups, and the robots classified within each group

| Group | Hum. | Sur. | Face | Body | Robots in the group |
|---|---|---|---|---|---|
| 1. Mechanical bots | 19.6 | 0.06 | 0.47 | 0.44 | Zenbo, Durus, Misty, Atlas, Ollie, Jibo, Cozmo, Cassie, Pr2, Walker, Loomo, Mip, Hitchbot, Handle (later), Spot, Handle (earier), Vector |
| 2. Androids | 49.3 | 0.18 | 0.81 | 0.9 | Valkyrie, Pepper, Asimo, Nao, Nexi, Robonaut, Icub, Cb2, Kojiro, Qrio |
| 3. Half-humanoids | 76.3 | 0.77 | 0.99 | 0.37 | Sophia, Bina48, Han |
| 4. Humanoids | 93.1 | 0.97 | 0.99 | 0.97 | Jiajia, Erica, Nadine |

of humanlikeness and subscales scores and the group assignment for particular robots.

In order to identify the language which people use to describe robots from the humanlikeness spectrum, I counted adjectives used for previously distinguished groups of robots. Adjectives are usually used for the description of the features and expressions of opinion in language [13,23], and are therefore good targets for attributes retrieval. Adjectives were counted for each robot and then normalized to the number of total words in the robot corpus in order to avoid enlarging the contribution of robots with more comments. For each group, I then counted the arithmetic mean of the frequencies of the words (due to a different number of robots in groups). Then, in order to identify adjectives that occur with unusual frequency in a given group, I counted scores for words according to the following equation:

$$score = \frac{f_w^2}{f_t},$$

where $f_w$ is the word frequency within the group, and $f_t$ is the word frequency in all groups. Such an estimation emphasizes words that are relatively frequent in a group in relation to other groups. The top 15 sorted adjectives with the highest score for each group are presented in Table 7.

Whereas for *mechanical bots* and *androids* groups it is hard to indicate any specific words (although the ones I identified all seem to be positively-valenced); for *half-humanoids* and *humanoids* it strikes that the words are related to the uncanny valley, i.e., 'scary' and 'creepy'; and also to the artificial-real dimension, i.e., 'human', 'real', 'fake', 'live', 'realistic', 'robotic', 'artificial' and 'android'. An extended list of relatively frequent adjectives is added to Supplementary Materials.

### 3.6 Uncanny Valley Awareness

I also wanted to test the awareness of UVH among commenters explicitly. I counted the frequencies of the 'uncanny

**Table 7** The most frequent words relative to other groups

| Mech. bots | Androids | Half-humanoids | Humanoids |
|---|---|---|---|
| Good | Human | Human | Human |
| Cool | Cool | Nuclear | Real |
| Awesome | Little | Real | Beautiful |
| Much | Good | Fake | Japanese |
| Great | Cute | Creepy | Chinese |
| Cute | Better | Scary | Robotic |
| New | Much | First | Female |
| Best | New | Sure | Realistic |
| Little | Awesome | Old | Artificial |
| Better | Great | Stupid | Scary |
| Next | Able | Last | Android |
| Happy | Real | God | Young |
| Big | Advanced | Live | Sexual |
| First | Amazing | Weird | Bad |
| Flat | Japanese | Realistic | Hot |

valley' term appearing in comments. For several robots, there were no occurrences of the term, meaning that the chi-square test could not be used to statistically test differences. I present the normalized frequency in Fig. 11. The plot shows that the occurrence of the 'uncanny valley' term is frequent for more humanlike robots, whereas its occurrence for other robots is none or near to none (with the exception of the Nexi robot). The public seems to be aware, at least to some extent, of the existence of the uncanny valley.

## 4 Discussion

The aims of this paper were as follows: to test the shape of the relationship between robot humanlikeness and sentiment scores, to examine which of the variables (eeriness, pleasantness, and attractiveness) are related to humanlikeness, to
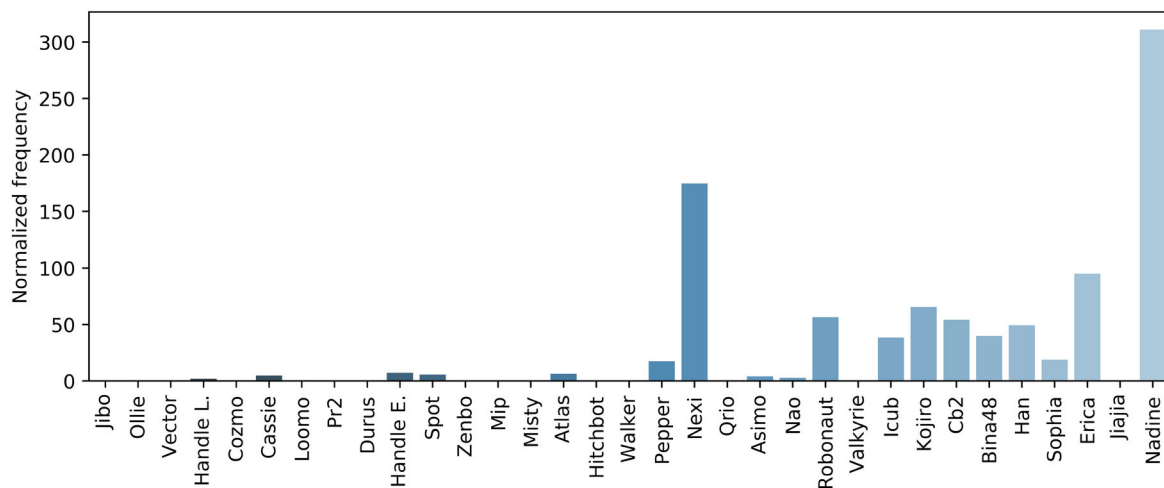
**Fig. 11** Normalized frequency of the 'uncanny valley' term occurrence; per 1000 comments. Robots are ordered by the humanlikeness score

test the impact of robot size on sentiment, and to characterize the specific, emotional words expressed toward robots. The study focused on providing ecologically-valid results about people's emotional reactions toward robots. The acquisition of comments from the YouTube video-sharing platform allowed relatively natural utterances, not affected by experimental conditions, to be examined.

The analysis of robot-related comments supports the presence of a specific attitude toward very humanlike robots called the uncanny valley. The results show that people use words relating to the concept of eeriness to describe very humanlike robots. Given the large sample of data (224,544 comments for 33 robots), this is strong evidence, that the uncanny valley is a real issue and doubts of its existence (e.g., [4,25]) are not valid. Emotions manifested in UVH are limited to eeriness. The shape of general humanlikeness and sentiment is linear – more humanlike robots elicit more negative sentiment. One of the subscales of humanlikeness – Facial Features – shows a non-linear relationship with eeriness and sentiment. The attractiveness, related to mate selection, is one of the explanations of the uncanny valley (e.g., [8,32]). My results show no relationship between attractiveness and humanlikeness, and do not support this explanation. Additionally, the study shows that the size of robots can influence the general emotions toward them (mediated by the perception of smaller robots as designed for play), which is in line with [61].

### 4.1 Shape of the Uncanny Valley

Mori [35] hypothesized the existence of a non-linear relationship between humanlikeness and affinity level. However, empirical studies advocate that affinity increases linearly across increasing humanlikeness [8,25]. The results of the YouTube comments analysis also support the hypothesis

that the relationship between humanlikeness and emotional valence (positive vs. negative) is linear, but not in the direction proposed by Kätsyri et al. [25]. According to the analysis, as humanlikeness increases, sentiment decreases. A reverse relationship seems to exist for eeriness—as humanlikeness increases, perceived eeriness increases.

As for factors that underlie the humanlikeness (according to Phillips et al. [40]), only the Body-Manipulators subscale (presence of hands, arms, torso, fingers, and legs) does not influence the sentiment score (nor the eeriness). The impact of this subscale for UVH is limited, which is interesting, as the Body-Manipulators subscale was previously found to be the greatest contributor to the humanlikeness of robots [40]. It seems that Surface Look (presence of eyelashes, head hair, skin, genderedness, nose, eyebrows, apparel) and Facial Features (presence of a face, eyes, head, mouth) are the most important humanlikeness subscales for the UVH. There was a linear relationship between Surface Look and sentiment, whereby higher levels of Surface Look were associated with more negative sentiment. With respect to eeriness, there was a positive relationship with Surface Look up to a certain degree, but at the highest levels of Surface Look the pattern reversed and eeriness perceptions decreased. The Facial Features subscale shows a sinusoidal pattern both for sentiment and eeriness—a very high score on this scale seems to greatly decrease sentiment and increase eeriness. This relationship seems to resemble the characteristic dip of UVH. The dimension of Facial Features reflects people's expectations that robots interact socially and effectively communicate with humans [40], therefore this is yet another argument for the involvement of social thinking in the uncanny valley effect [34,58].

Mathur and Reichling [34] showed a cubic relationship between mechano-humanness and likability after conducting an experimental study. Because they used images of

robot faces as stimuli, one may have suspected that the relationship would be similar to the relationship between Facial Features and sentiment in this study. However, cubic functions were mirrored (in [34], the cubic function approaches positive infinity, and in my study the function approaches negative infinity). Mathur and Reichling [34] asked subjects to estimate friendliness versus creepiness of possible interactions with a robot after a static image display. This stands in contrast to motion picture stimuli from my study. The method of stimuli presentation (images vs. movies) influence the elicitation of emotions [12] and this may be the reason for the different shapes of the obtained models. Possibly, people watching robot movies are able to judge the social behavior of robots, attributing agency, and experience (see [16]) far better than while watching images. The usage of static images in Mathur and Reichling [34] puts emphasis on visual aspects of robots, thus variable of static images may be more similar to the influence of surface look. This is in line with the similarity between results of eeriness index and Surface Look subscale relationship (see Fig. 6) and mechano-humanness and likability relationship from Mathur and Reichling [34]. Many of the uncanny valley studies focus mainly on visual aspects of robots, but the results of Stein and Ohler [50] showed that interacting characters may be seen as more eerie, depending on the beliefs of observers—if they think that characters are controlled by artificial intelligence, they assess characters as more eerie than when thinking that they are controlled by a human. This suggests that the Theory of Mind (e.g., [11]) factor may be involved in the uncanny valley effect. Mori [35] proposed that the movement of robots exaggerates the eeriness, but this has not been confirmed [41]. Presumably, this hypothesis should be modified, i.e., not the simple movement (as in the study of [41], which was a simple door-knocking movement), but the complex movement of robots, which is perceived as specific behavior, exaggerates or even changes the effect. This would suggest the necessity of analyzing the impact of robots' behavior in the context of UVH.

Considering the comparison of models and the plots of relationship, it seems that the rightmost part of Mori's plot (the "leave" from the valley of the most humanlike robots) is questionable. Under conditions taking into account the mind attribution, the relationship resembles a cliff (as suggested by [3]) rather than a valley. Either the shape of UVH should be reconsidered, or perhaps it is impossible to determine the real shape of the uncanny valley as there are no real robots that are indistinguishable or nearly indistinguishable from human beings, due to the current state of technology.

Based on the obtained results and the above-mentioned literature, a revised version of the uncanny valley plot and its modifications is considered in Fig. 12.
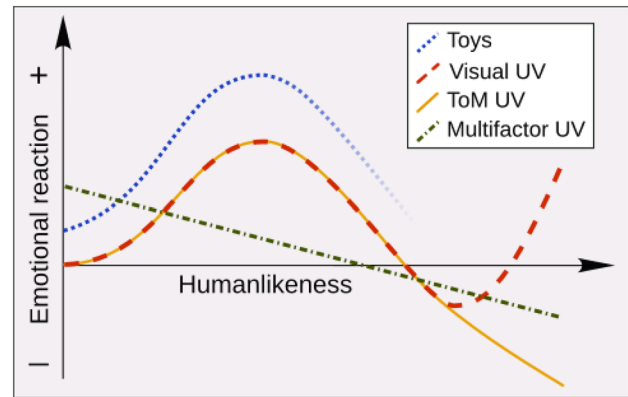


**Fig. 12** Modifications of UV relationship; Toys—agents perceived as toys, visual UV—relationship between humanlikeness based on visual aspects and sentiment, ToM UV—relationship between humanlikeness based on attribution of human mind and sentiment, multifactor UV—relationship between multifactorial humanlikeness and sentiment. Detailed discussion in the text

### 4.2 Emotions Describing UVH

The regression analysis showed that eeriness, pleasantness, and attractiveness were not equally related to humanlikeness (H2), and in fact, only the eeriness index is associated with humanlikeness. While controlling eeriness, pleasantness (defined by the property of "giving pleasure or contentment to the mind or senses"), and attractiveness (defined by the property of "being very pleasing to look at"[13]) did not emerge as significant variables for explaining the uncanny valley effect.

The exploratory analysis of adjectives reflects the attitudes evident in the regression analysis. For less humanlike groups (mechanical bots and androids), relatively the most frequent adjectives were positive or neutral and not specific. For the most humanlike robots, relatively the most frequent adjectives are related to the perception of eeriness, i.e., 'scary' and 'creepy', and also for artificial-real dimension, i.e., 'human', 'real', 'fake', 'live', 'realistic', 'robotic', 'artificial' and 'android'. This means that the uncanny valley feelings, as well as humanlikeness itself, are among the most discussed topics in the comments specific to very humanlike robots. The emotional adjectives from the list seem to be the most suitable words for measuring a self-reported decrease in affinity related to the uncanny valley. This addresses the suggestion of Kätsyri et al. [25] about the necessity of such empirical studies. For humanoids, two distinguished words ('sexual' and 'hot') may be related to robots' sexualization phenomenon identified by Strait et al. [52]. This observation supports their claim that objectification of female-gendered robots is a real issue.

---

[13] Both definitions are taken from https://www.merriam-webster.com/thesaurus/ [Accessed: 07.12.2020].

It is also worth mentioning that some people are aware of the uncanny valley effect, which is shown in Fig. 11, and this topic is popular among the internet community. This part of the commenters either know the definition of UVH or implicitly understand it, because occurrence of the term is limited to humanlike robots (not generally to robots). The case of Nexi robot, which had relatively more mentions of the 'uncanny valley' term than robots with similar humanlikeness, suggests the implicit understanding of the phenomenon. The Nexi robot has highly developed facial expressions, which also suggests the importance of mind attribution for UVH. Perhaps, in future experiments, the participants' awareness of the uncanny valley should be controlled in order not to allow the results of self-report experiments to be biased by implicit knowledge.

## 4.3 Impact of Robots' Size

The results show a significant relationship between the height of robots and sentiment scores. Additionally, perception of robots as more playable (as toys) is a mediator of this relationship.

Mäkäräinen et al. [33] came up with the concept of 'funcanny valley', i.e., artificial characters may be seen in a funny way, regardless of their uncanniness. They conducted a study using, among others, exaggerated smiling characters which elicited a positive reaction, despite their increasing strangeness. They suggested that the negative affective reaction described by the uncanny valley concept could, in some cases, evoke the sensation of amusement, funniness, and humorousness. Although their study had a different methodology and focused on human characters, the operationalization of the funcanny valley concept to my results may identify the factor of size as a variable in the funcanny valley. However, this interpretation does not explain the results of Mäkäräinen et al. [33], and further analyses are needed to identify how broad this concept is.

Perhaps the perception of artificial/robotic characters as created for amusement masks the uncanny effect, which may explain some differences in the assessments of uncanny characters.

## 4.4 Limitations

The study presented in this paper has some limitations that should be highlighted. Firstly, demographic information of commenters is not available to acquire on YouTube. Therefore, it is not clear if the acquired data is representative for the population. The sample selection might be biased by the YouTube algorithm, which may recommend watching robot videos to people interested in the topic. As a result, people who watch robot videos may be exposed to more robot videos. This is a valid issue because seeing more films

portraying robots tends to be associated with more positive attitudes toward robots [44]. Also, not all users are willing to put a comment under videos. Type of personality or attitudes toward internet media may influence whether or not someone shares their opinion online (see [38]). However, one may notice that type of personality or prior experience in research participation may influence the willingness to participate in laboratory studies either (see [46]). Some people may not want to participate in scientific studies but are willing to express their opinion on the Internet. It has been shown that internet comments analysis may provide valuable information about attitudes and can help to understand human behavior (e.g., [5,24]), therefore such an analysis, despite its weakness, may be beneficial for HRI research.

Additionally, the findings are based on English comments searched for the US region. Given that cultural factors might influence attitudes toward social robots and the way we respond to them [28], the findings should be interpreted with caution due to the potential generalisability issue. It would be valuable to conduct a similar analysis for other languages and regions in the future.

Furthermore, the context and narratives in which robots are presented may have an impact on the viewer's sentiment. For example, the word 'eerie' may not always reflect the sentiment toward the robot but could refer to other things presented in the video. I utilized a few means to decrease the possibility of such confounds. Firstly, I limited video search only to short videos below 4 min. The longer video is, the greater possibility that it will contain unwanted narratives or other unexpected content. Secondly, I excluded responses to the main comments (sub-comments), as they may explore side topics. Thirdly, I used multiple numbers of videos for each robot (45.9 on average, see the "Appendix" for individual numbers), therefore the effect of video context presumably has been averaged.

Although the initial number of robots prepared for the analysis was big (246), after filtering the sample size decreased to 33. This number is derived from the popularity of robots on the Internet and possibly, as robots are seen as an increasingly interesting topic, it will become possible to conduct similar analyses with a bigger sample in the future. The cut-off number of comments for robots (more than 200) was a trade-off between preserving the initial sample size and not leaving robots with too few comments for unbiased analysis. Small corpora (with fewer than 200 comments) may show random results, due to drift to various topics of movies.

In the analysis of sentiment and emotional indicators, negative forms of phrases have not been taken into account. During the preparation of the analysis, I made a few tries with negation detection in YouTube comments—-I tried to negate all the words in a sentence when 'not' occurs. However, the lack of punctuation in many comments caused problems. E.g., the algorithm negated all the words in comments with

multiple sentences without punctuation, which distorted the results. As Heerschop et al. [19] showed that simple inversion of the polarity of sentiment when negation occurs had a marginal effect on the improvement of performance (even for more structured text than internet comments), I used the conventional method of frequency counting with *AFINN* package without negation handling.

The values of $R^2$ in models analysis (Sects. 3.1, 3.2) are relatively low (besides the relationships with Facial Features), which may be justified by the various topics of analyzed movies. However, a large number of comments should compensate for this randomness and reveal underlying attitudes toward robots.

### 4.5 Future Work

The YouTube comments seem to be a rich source of information regarding attitudes toward robots. Despite its limitations, future HRI studies analyzing internet comments may provide more explanatory insights due to the advantages of ecological validity. Besides the possible sentiment analysis of comments in languages other than English, deeper semantic analyses may provide valuable information about what influences the acceptance of robots. Also, methods of Natural Language Processing may help understand the causes of the uncanny valley.

## Appendix

See Table 8.

**Table 8** List of all analyzed robots with number of comments and videos, means and standard deviations for sentiment, eeriness, pleasantness, attractiveness, humanlikeness, Surface Look, Facial Features, Body Manipulators scores, and height [cm]

| Name | Com. no. | Vid. no. | M sent. | SD sent. | Eerie. | Pleas. | Attr. | HL | Surf. | Face | BM | Height |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Asimo | 7090 | 167 | 0.6 | 3.01 | 1.2 | 5.7 | 7.8 | 45.4 | 0.0 | 0.5 | 1.0 | 130.0 |
| Atlas | 85507 | 134 | −0.09 | 3.18 | 1.4 | 6.1 | 6.6 | 31.0 | 0.1 | 0.4 | 0.6 | 188.0 |
| Bina48 | 751 | 5 | −0.12 | 3.06 | 8.0 | 6.3 | 9.5 | 73.0 | 0.9 | 1.0 | 0.0 | 44.2 |
| Cassie | 2082 | 15 | 0.6 | 2.69 | 1.9 | 6.0 | 6.2 | 14.9 | 0.0 | 0.0 | 0.2 | 100.0 |
| Cb2 | 552 | 11 | −0.22 | 2.41 | 9.7 | 2.9 | 6.5 | 64.5 | 0.3 | 1.0 | 1.0 | 130.0 |
| Cozmo | 12588 | 167 | 1.7 | 3.22 | 0.3 | 7.7 | 15.0 | 13.9 | 0.1 | 0.8 | 0.2 | 6.4 |
| Durus | 219 | 2 | 0.44 | 3.0 | 1.2 | 9.0 | 9.6 | 21.0 | 0.0 | 0.2 | 0.4 | 180.0 |
| Erica | 736 | 11 | 0.49 | 3.66 | 3.6 | 7.2 | 9.3 | 89.6 | 0.9 | 1.0 | 0.9 | 166.0 |
| Han | 202 | 7 | 0.29 | 3.38 | 8.1 | 4.0 | 5.8 | 77.0 | 0.7 | 1.0 | 0.3 | 78.3 |
| Handle E. | 9658 | 25 | 0.18 | 3.0 | 2.4 | 6.1 | 6.4 | 21.2 | 0.0 | 0.2 | 0.6 | 198.0 |
| Handle L. | 5197 | 9 | 0.36 | 2.83 | 2.2 | 6.6 | 7.9 | 8.8 | 0.0 | 0.0 | 0.6 | 200.0 |
| Hitchbot | 3466 | 19 | 0.36 | 3.51 | 0.4 | 8.2 | 6.8 | 31.6 | 0.2 | 0.9 | 0.9 | 120.0 |
| Icub | 520 | 46 | 0.69 | 3.2 | 3.0 | 8.1 | 7.4 | 55.4 | 0.2 | 0.8 | 1.0 | 104.0 |
| Jiajia | 308 | 13 | 0.18 | 3.01 | 2.4 | 4.7 | 10.6 | 93.3 | 1.0 | 1.0 | 1.0 | 160.0 |
| Nao | 3609 | 19 | 0.61 | 2.9 | 1.4 | 6.4 | 9.9 | 45.9 | 0.0 | 0.9 | 1.0 | 57.4 |
| Nexi | 858 | 3 | 0.07 | 3.47 | 11.6 | 9.0 | 8.7 | 42.3 | 0.3 | 1.0 | 0.6 | 122.0 |
| Ollie | 2653 | 39 | 0.76 | 2.55 | 0.2 | 6.3 | 7.0 | 3.6 | 0.0 | 0.1 | 0.1 | 12.0 |
| Pepper | 1140 | 62 | 0.9 | 3.03 | 2.3 | 5.3 | 11.2 | 40.5 | 0.2 | 1.0 | 0.9 | 120.0 |
| Pr2 | 269 | 17 | 1.45 | 2.87 | 0.0 | 9.4 | 9.9 | 20.0 | 0.1 | 0.7 | 0.8 | 149.0 |
| Qrio | 297 | 18 | 0.8 | 3.52 | 0.6 | 2.7 | 6.4 | 44.1 | 0.1 | 0.7 | 0.8 | 60.0 |
| Robonaut | 354 | 37 | 0.71 | 3.41 | 1.6 | 5.7 | 5.9 | 47.0 | 0.1 | 0.5 | 1.0 | 101.6 |
| Sophia | 7925 | 89 | −0.04 | 3.32 | 5.8 | 5.7 | 6.6 | 78.9 | 0.7 | 1.0 | 0.8 | 167.0 |
| Spot | 68263 | 93 | 0.31 | 4.84 | 3.4 | 6.6 | 8.8 | 24.3 | 0.0 | 0.0 | 0.4 | 84.0 |
| Valkyrie | 839 | 36 | 1.02 | 3.18 | 0.3 | 6.7 | 5.3 | 50.6 | 0.2 | 0.8 | 1.0 | 190.0 |
| Vector | 6373 | 241 | 0.9 | 2.94 | 0.7 | 6.3 | 10.8 | 6.1 | 0.0 | 0.8 | 0.2 | 6.1 |
| Walker | 287 | 11 | 1.16 | 3.1 | 2.7 | 7.3 | 7.3 | 33.9 | 0.0 | 0.5 | 0.3 | 145.0 |
| Zenbo | 320 | 28 | 0.82 | 2.97 | 4.0 | 7.9 | 11.9 | 24.9 | 0.2 | 0.9 | 0.1 | 62.0 |

## Declarations

## References

1. Ahmed K, El Tazi N, Hossny AH (2015) Sentiment analysis over social networks: an overview. In: 2015 IEEE international conference on systems, man, and cybernetics. IEEE, pp 2174–2179
2. Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51(6):1173
3. Bartneck C, Kanda T, Ishiguro H, Hagita N (2007) Is the uncanny valley an uncanny cliff? In: Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on, pp 368–373
4. Bartneck C, Kanda T, Ishiguro H, Hagita N (2009) 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. https://doi.org/10. 1109/ROMAN.2009.5326351
5. Beigi G, Hu X, Maciejewski R, Liu H (2016) An overview of sentiment analysis in social media and its applications in disaster relief. In: Sentiment analysis and ontology engineering, Springer, pp 313–340
6. Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93(3):491–507
7. Burleigh TJ, Schoenherr JR (2015) A reappraisal of the uncanny valley: categorical perception or frequency-based sensitization? Front Psychol 5:1488
8. Burleigh TJ, Schoenherr JR, Lacroix GL (2013) Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. Comput Hum Behav 29(3):759–771
9. Burnham KP, Anderson DR (2002) A practical information-theoretic approach. Model selection and multimodel inference, 2nd edn. Springer, New York, p 2
10. Cheetham M, Wu L, Pauli P, Jancke L (2015) Arousal, valence, and the uncanny valley: Psychophysiological and self-report findings. Front Psychol 6:981
11. Dennett DC (1971) Intentional systems. J Philos 68(4):87–106
12. Detenber BH, Simons RF, Bennett GG Jr (1998) Roll 'em!: the effects of picture motion on emotional responses. J Broadcast Electron Media 42(1):113–127
13. Fei G, Liu B, Hsu M, Castellanos M, Ghosh R (2012) A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. In: Proceedings of COLING 2012: Posters, pp 309–318
14. Fernández-Berrocal P, Extremera N (2006) Emotional intelligence and emotional reactivity and recovery in laboratory context. Psicothema 18:72–78
15. Friedman B, Kahn Jr PH, Hagman J (2003) Hardware companions? What online aibo discussion forums reveal about the human–robotic relationship. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp 273–280
16. Gray K, Wegner DM (2012) Feeling robots and human zombies: mind perception and the uncanny valley. Cognition 125(1):125–130
17. Guzman E, Azócar D, Li Y (2014) Sentiment analysis of commit comments in github: an empirical study. In: Proceedings of the 11th working conference on mining software repositories, pp 352–355
18. Hajar M et al (2016) Using youtube comments for text-based emotion recognition. Procedia Comput Sci 83:292–299
19. Heerschop B, van Iterson P, Hogenboom A, Frasincar F, Kaymak U (2011) Analyzing sentiment in a large set of web data while accounting for negation. In: Advances in intelligent web mastering, vol 3, Springer, pp 195–205
20. Ho CC, MacDorman KF (2010) Revisiting the uncanny valley theory: developing and validating an alternative to the godspeed indices. Comput Hum Behav 26(6):1508–1518
21. Ho CC, MacDorman KF, Pramono ZD (2008) Human emotion and the uncanny valley: a glm, mds, and isomap analysis of robot video ratings. In: 2008 3rd ACM/IEEE international conference on human–robot interaction (HRI), pp 169–176
22. Hover QR, Velner E, Beelen T, Boon M, Truong KP (2021) Uncanny, sexy, and threatening robots: The online community's attitude to and perceptions of robots varying in humanlikeness and gender. In: Proceedings of the 2021 ACM/IEEE international conference on human–robot interaction, pp 119–128
23. Hu M, Liu B (2004) Mining opinion features in customer reviews. In: Proceedings of 19th national conference on artificial intelligence (AAAI'04), vol 4, pp 755–760
24. Jelodar H, Wang Y, Orji R, Huang S (2020) Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. IEEE J Biomed Health Inform 24(10):2733–2742
25. Kätsyri J, Förger K, Mäkäräinen M, Takala T (2015) A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Front Psychol 6:390
26. Kätsyri J, de Gelder B, Takala T (2019) Virtual faces evoke only a weak uncanny valley effect: an empirical investigation with controlled virtual face images. Perception 48(10):968–991
27. Kim SY, Schmitt BH, Thalmann NM (2019) Eliza in the uncanny valley: anthropomorphizing consumer robots increases their perceived warmth but decreases liking. Mark Lett 30(1):1–12
28. Lim V, Rooksby M, Cross ES (2021) Social robots on a global stage: establishing a role for culture during human–robot interaction. Int J Soc Robot 13(6):1307–1333
29. Löffler D, Dörrenbächer J, Hassenzahl M (2020) The uncanny valley effect in zoomorphic robots: the u-shaped relation between animal likeness and likeability. In: Proceedings of the 2020

ACM/IEEE international conference on human–robot interaction, pp 261–270

30. Łupkowski P, Rybka M, Dziedzic D, Włodarczyk W (2019) The background context condition for the uncanny valley hypothesis. Int J Soc Robot 11(1):25–33

31. MacDorman KF, Chattopadhyay D (2016) Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition 146:190–205

32. MacDorman KF, Green RD, Ho CC, Koch CT (2009) Too real for comfort? Uncanny responses to computer generated faces. Comput Hum Behav 25(3):695–710

33. Mäkäräinen M, Kätsyri J, Förger K, Takala T (2015) The funcanny valley: a study of positive emotional reactions to strangeness. In: Proceedings of the 19th international academic mindtrek conference. ACM, pp 175–181

34. Mathur MB, Reichling DB (2016) Navigating a social world with robot partners: a quantitative cartography of the uncanny valley. Cognition 146:22–32

35. Mori M (1970) The uncanny valley. Energy 7(4):33–35

36. Ngai EW, Tao SS, Moon KK (2015) Social media research: theories, constructs, and conceptual frameworks. Int J Inf Manag 35(1):33–44

37. Nielsen FÅ (2011) A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 workshop on making sense of microposts (2011)

38. Orchard LJ, Fullwood C (2010) Current perspectives on personality and internet use. Soc Sci Comput Rev 28(2):155–169

39. Palomäki J, Kunnari A, Drosinou M, Koverola M, Lehtonen N, Halonen J, Repo M, Laakasuo M (2018) Evaluating the replicability of the uncanny valley effect. Heliyon 4(11):e00939

40. Phillips E, Zhao X, Ullman D, Malle BF (2018) 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

41. Piwek L, McKay LS, Pollick FE (2014) Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. Cognition 130(3):271–277

42. R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

43. Ratajczyk D, Jukiewicz M, Lupkowski P (2019) Evaluation of the uncanny valley hypothesis based on declared emotional response and psychophysiological reaction. Bio Algorithms Med Syst. https://doi.org/10.1515/bams-2019-0008

44. Riek LD, Adams A, Robinson P (2011) Exposure to cinematic depictions of robots and attitudes towards them. In: Proceedings of international conference on human–robot interaction, workshop on expectations and intuitive human–robot interaction, Citeseer, vol 6

45. Rosenthal-Von Der Pütten AM, Krämer NC (2014) How design characteristics of robots determine evaluation and uncanny valley related responses. Comput Hum Behav 36:422–439

46. Saliba A, Ostojic P (2014) Personality and participation: who volunteers to participate in studies. Psychology 5, 230–243

47. Seyama J, Nagayama RS (2007) The uncanny valley: effect of realism on the impression of artificial human faces. Presence Teleoperators Virtual Environ 16(4):337–351

48. Shank DB, Graves C, Gott A, Gamez P, Rodriguez S (2019) Feeling our way to machine minds: people's emotions when perceiving mind in artificial intelligence. Comput Hum Behav 98:256–266. https://doi.org/10.1016/j.chb.2019.04.001

49. Siersdorfer S, Chelaru S, Nejdl W, San Pedro J (2010) How useful are your comments? Analyzing and predicting youtube comments and comment ratings. In: Proceedings of the 19th international conference on World wide web. ACM, pp 891–900

50. Stein JP, Ohler P (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 160:43–50

51. Strait M, Ramos AS, Contreras V, Garcia N (2018) Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. In: 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, pp 452–457

52. Strait MK, Aguillon C, Contreras V, Garcia N (2017) The public's perception of humanlike robots: Online social commentary reflects an appearance-based uncanny valley, a general fear of a "technology takeover", and the unabashed sexualization of female-gendered robots. In: 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 1418–1423

53. Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29(1):24–54

54. Tay T, Low R, Loke H, Chua YL, Goh Y (2018) Uncanny valley: A preliminary study on the acceptance of Malaysian urban and rural population toward different types of robotic faces. In: IOP conference series: materials science and engineering, vol 344. IOP Publishing, p 012012. https://doi.org/10.1088/1757-899x/344/1/012012

55. Thelwall M (2018) Social media analytics for youtube comments: potential and limitations. Int J Soc Res Methodol 21(3):303–316

56. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014) mediation: R package for causal mediation analysis. J Stat Softw 59(5):1–38

57. Urgen BA, Li AX, Berka C, Kutas M, Ishiguro H, Saygin AP (2015) Predictive coding and the uncanny valley hypothesis: evidence from electrical brain activity. In: Cognition: a bridge between robotics and interaction, pp 15–21

58. Urgen BA, Kutas M, Saygin AP (2018) Uncanny valley as a window into predictive processing in the social brain. Neuropsychologia 114:181–185

59. Vlachos E, Tan ZH (2018) Public perception of android robots: indications from an analysis of youtube comments. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 1255–1260

60. Wagner D, Billinghurst M, Schmalstieg D (2006) How real should virtual characters be? In: Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology. ACM, p 57

61. Wang I, Smith J, Ruiz J (2019) Exploring virtual agents for augmented reality. In: Proceedings of the 2019 CHI conference on human factors in computing systems. ACM, p 281

62. Wang S, Lilienfeld SO, Rochat P (2015) The uncanny valley: existence and explanations. Rev Gen Psychol 19(4):393–407

63. Yu CE (2020) Humanlike robots as employees in the hotel industry: thematic content analysis of online reviews. J Hosp Mark Manag 29(1):22–38

**Dawid Ratajczyk** is a cognitive scientist and a PhD student at the Adam Mickiewicz University in Poznan, Poland. Besides a master's degree in Cognitive Science, he also received a bachelor's degree in Biology. His current research interest is focused on social robotics, human-computer interaction, virtual reality, and language processing.