Human TECHNOLOGY

ISSN: 1795-6889

https://ht.csr-pub.eu

Volume 19(1), May 2023, 62-81

# TOWARDS THE STANDARDISED SET OF STIMULI FOR THE UNCANNY VALLEY PHENOMENON STUDIES

Paweł Łupkowski
*Adam Mickiewicz University*
*Poland*
*ORCID 0000-0002-5335-2988*

Dawid Ratajczyk
*Adam Mickiewicz University*
*Poland*
*ORCID 0000-0002-5335-2988*

Abstract: *This paper presents a pre-validated set of 12 static 3D stimuli for the needs of uncanny valley-related studies. We provide the set along with guidelines on how to use it, which are based on the aggregated data analysis covering previous laboratory and online studies. Guidelines cover the models' characteristics (in terms of human-likeness and other visual traits), issues related to the stimuli presentation and study groups. As the set is publicly available, we believe that it enhances future studies aimed at replicating the uncanny valley effect.*

Keywords: *uncanny valley, visual stimuli, replicability, human-robot interaction.*


DOI: https://doi.org/10.14254/1795-6889.2023.19-1.5

## INTRODUCTION

As Palomaki et al. (2018, p. 3-4) point out, the International Federation of Robotics estimated that by 2019 more than 42 million robots had been sold for personal use; meaning, they were quickly becoming an unavoidable element of our social ecosystem. This ecosystem consists of more and more robots with human-like traits, (i.e., social robots and various humanoid robots). What is important to understand about this ecosystem is that it should be perceived widely, not only as inhabited by real physical robots, but also their digital counterparts: computer graphics in video games, virtual assistants, virtual and augmented reality agents.

Picarra, Giger, Pochwatko, and Mozaryn (2016, p. 18) notice that a growing number of such robots will have an unavoidable effect on our expectations concerning robots: "(...) designing robots with human-like traits can enhance their interactive and social proficiency. Also, different degrees of human-likeness seem to impact differently potential user's expectations and behavior". This brings us to the idea of the so-called uncanny valley hypothesis (hereafter UVH). The hypothesis was formulated by Masahiro Mori (see Mori, MacDorman, and Kageki 2012). Mori hypothesises that when we present a subject with a series of different human-like models (including robots), certain models will trigger negative reactions (uneasiness, eeriness). As he suggests, this uneasiness will be caused by the almost humanlike characteristics of robots. We may imagine models presented in order, from the least human-like (like e.g., robotic arm) to the most human-like ones on the X axis. On the Y axis, we could present the affinity level. The specific name of the dependent variable is still under debate (e.g., Kätsyri, Förger, Mäkäräinen, & Takala, 2015; see also Ratajczyk, 2022). It may be referred to as affinity, likability, general emotional reaction (positive/negative) and others. According to Mori's suggestion, we will observe a growing level of affinity as we move towards human-like models, but past a certain level of human-likeness, the level of affinity will rapidly decrease. This is the "valley" (or as MacDorman and Norri Kageki call it "descent into eeriness" – Mori et al. 2012).

Recently, we have observed a growing body of UVH studies. However, as we may read in the extensive meta-analysis of such studies, "[I]t is surprising that empirical evidence for the uncanny valley hypothesis is still ambiguous if not non-existent." (Katsyri et al., 2015). One of the possible causes of such a situation, identified by Katsyri et al., is the stimuli used for the research. "Mori himself used anecdotal examples to characterise different degrees of human-likeness (industrial robots, toy robots, prosthetics and even puppets, corpses and zombies)" (Katsyri et al., 2015). In the literature, we can observe several different approaches to the study of UVH. Some researchers rely on real robots, but the vast majority of studies use various visual stimuli, such as 2D static pictures, animations, or videos. It is very seldom met that one set of stimuli appears in more than one study. This is strongly pointed out in Palomaki et al. (2018). Based on their attempts of replicating UVH effects, these authors conclude the following: "[t]o our knowledge there is generally a lack of pre-validated stimulus materials known to reliably, robustly, and repeatedly elicit the UV effect, such as a common repository of "uncanny" robot pictures (or other objects) – or specific guidelines to create such stimuli" (Palomaki et al., 2018).

This quote provides a direct motivation for this paper. As we have conducted four studies in which we have used the same set of visual stimuli, we have decided to analyse the

aggregated data in order to learn more about the models as well as produce the guideline for their use for future studies.

In what follows, we discuss related work and provide a background for certain design decisions and choices related to the set of 12 stimuli we use. After this, we introduce the set of models in section "The set". Next, we shortly describe four studies in which these models were used and present analyses of the aggregated data taken from these studies. In our analyses we address the following questions:

(i) Do the online and laboratory studies have different results when it comes to human-likeness evaluation and emotional reaction?

(ii) Do we observe gender differences in the models' assessment?

(iii) Is there a correlation between the evaluation of human-likeness and time needed for the evaluation?

Section "ABOT Database Robot Human-Likeness Predictor Study" presents the evaluation of our data set in terms of human-likeness by using the ABOT Database Predictor tool. Section "Selected studies and their implications for the stimuli usage" contains selected results from the studies with the set, which offer potentially useful implications for the future use of the set. We end the paper with a discussion of the limitations of our study and the guideline for the future usages of the stimuli set. The guideline covers: models' characteristics, stimuli presentation and study groups recommendations.

It is worth stressing that in this paper we do not focus on the examination of the uncanny valley effect. Our main goal is to test additional variables of a potentially disturbing nature (assessment of humanlikeness, experimental conditions), within the set of models which was used in our previous studies. Hence, we have decided to use the aggregated data from all of the studies with the described models. This meta-analysis allows for pointing out certain consistent observations across different groups of subjects (overall 237 participants) and different experimental conditions (laboratory and online). The main contribution here is the pre-validated set of models (publicly available for researchers) along with the guideline for its use and detailed discussion of factors and variables which should be controlled in future studies with the described set. As such we believe that this work enhances future studies aimed at replicating the uncanny valley effect (for computer generated graphics).

## RELATED WORK

There are various sets of pictures of robots which were used as stimuli in UVH studies and are available for researchers. Mathur et al. (2020) shared a set of pictures of robots and humans acquired *via* Internet search with the strict exclusion criteria. In their study, they tested one of the UVH explanations (namely the categorization difficulty) with the use of this set. Narrower version of the set (only robots, without pictures of actual human faces) was also used by Mathur and Reichling (2016) in an attempt to replicate the shape of the uncanny valley proposed by Mori. The set of pictures used in these studies consists only of head views of real robots and human faces. Models are presented on different backgrounds and they differ slightly when it comes to their presentation (head position, neck visibility, light level). In the case of our set of stimuli we have decided on the evaluation of CGI (computer generated imagery) models. The reason for that was that the uncanny valley effect is also

recognized in animations and that we planned to use our models mainly for on-screen presentations. Another difference is that in our set models are presented with the whole body and are all in neutral pose. The use of CGI allows us to prepare the scene and the pose of a model freely, according to the needs of a particular study. Also, as the models are rendered in the 3D we can easily modify them and export only selected parts (e.g. only faces or only hands, etc.).

There are several studies which evaluated CGI characters from popular cartoons and animated movies (e.g., Dill et al., 2012; Katsyri, Makarainen, & Takala, 2017). The results show that semi-realistic movies more likely elicit uncanny feeling contrary to cartoonish and fully realistic movies. Mentioned studies used well known, popular characters from movies. Because cultural connotations may influence the emotions that characters evoke (for a discussion of such connotations related to the presented set of models see Łupkowski & Gierszewska, 2019) it is important to build additional knowledge on the basis of the evaluation of neutral characters and draw conclusions for future UVH research. Thus our decision was to prepare new models for the set and not use already known characters.

There are also studies which focused on identification of visual factors of characters which may cause specific emotional responses in the audience. Schwind, Wolf, and Henze (2018) presented design guidelines based on UVH research for CGI developers to prevent the uncanny valley. Zhang et al. (2020) prepared similar general guidelines for UVH researchers in order to reduce inconsistencies among studies. McDonnell, Breidt, and Bulthoff (2012) tested the influence of rendering styles on virtual agents' reception. They found that particular style of rendering may change the pleasantness of characters, but it does not change the trust toward these agents. All three studies focused on visual aspects of stimuli. We also discuss these issues for our set of models.

Additionally, Zhang et al. (2020) present the results of the UVH literature review, where they juxtapose online and laboratory studies, but they did not compare the differences explicitly. This is the topic we analyse for our set of models as we think that it is important to know whether the same model is evaluated differently in an online and in a laboratory study.

## THE SET

Our set of models prepared for the needs of UVH studies described below consists of 12 static 3D images. As such, these studies are in line with an approach to study UVH by presenting static pictures to the subjects (see e.g., MacDorman, Green, Ho, & Koch, 2009; Geller, 2008; Piwek, McKay, & Pollick, 2014 or Schneider, Wang, & Yang, 2007).

The models (which can be found in the Appendix) were retrieved from 3D character banks (see http://tf3dm.com and https://www.mixamo.com/) and initially rendered using the Unity environment (https://unity3d.com/). All the models were chosen arbitrarily and then consulted with two designers experienced in game development. The models can be used in 2D studies (as pictures), as well as in static and dynamic (after applying an animation) 3D studies.

The models cover various levels of degrees of human-likeness (DOH), ranging from simplistic robots, through androids, animated characters, and up to human models. As we have discussed in (Łupkowski, Rybka, Dziedzic, & Włodarczyk, 2017), for the initial tests

we have also included three models presenting zombies as the visual stimuli. Based on the evaluation of the human-likeness issues for these models reported in (Łupkowski et al., 2017, p. 130) and the subject literature, like the aforementioned analysis of (Katsyri et al., 2017), we have resigned from further use of these models. The key features which we took into account were: visible facial features, detail level of the model (e.g., visible hands and fingers), and overall style of the model. For example, models intended as simplified robots do not have visible eyes and their hands and feet are not detailed. Also, the joints of various body parts are clearly visible, which gives them a very mechanical appearance. Models have the same height and are presented with a front face in a neutral pose. This presentation allows evaluation of body proportions, body elements (such as hands, feet), and even facial expressions.

It is worth noting that the set intentionally does not cover photo-realistic human renders. For all the models, we are dealing with computer-generated graphics only in order to have a consistent appearance. As pointed out by Zibrek & McDonnell (2014) different styles of rendering may possibly introduce additional confounding variables.

Along with this paper, we share the complete set of models. It may be downloaded from: https://osf.io/sxmju/?view_only=c7d7d042b5d2440aa3f66e6fb02fa538. This package contains all the models in the 3D *.blend format. For the needs of this paper we have re-rendered all the models with the open-source Blender 3D software (https://www.blender.org/). The aim was to make potential stimuli modifications easier and accessible for a wide audience as Blender 3D files allow for: (i) easy and fast rendering of 2D files of high resolution; (ii) easy modifications of the models (perspective change or lights placement on the scene); (iii) potential application of the models in different study scenarios (animation, virtual reality, augmented reality). This allows for easy preparation of the stimuli for the needs of future studies.

## ANALYSES OF THE AGGREGATED RESULTS

### Studies Selected for the Analysis

For the need of analysis presented below we aggregate data retrieved from the studies listed below. In all of them the same set of visual stimuli was used.

1. Łupkowski, Rybka, Dziedzic, & Włodarczyk *(2019), The Background Context Condition for the Uncanny Valley Hypothesis;*
2. Łupkowski, Gierszewska (2019), *Attitude Towards Humanoid Robots and the Uncanny Valley Hypothesis;*
3. Ratajczyk, Jukiewicz, & Łupkowski (2019), *Evaluation of the uncanny valley hypothesis based on declared emotional response and psychophysio-logical reaction;*
4. Łupkowski, Kulinski (2020), Aspekty percepcji wirtualnych modeli humanoidalnych w kontekście zjawiska doliny niesamowitości *[Perceiving human-like humanoid models – case study for the uncanny valley studies].*

In these studies models were presented one by one to our subjects (in different orders) along with questions. These studies had different aims and research questions, but all four shared the same question addressing the presented models' degree of human-likeness. The subjects performed this evaluation by answering the following question: *How much does the*

*presented model resemble a human?* (Answer on a scale 1-5, where: (1) Completely not human-like; (2) Rather not human-like; (3) Beginning to look human-like; (4) Rather human-like; (5) Completely human-like).

Three of the studies also asked the question addressing the comfort dimension: *How comfortable are you when you watch this model in a given environment?* (Answer on a scale 1-5, where: (1) very comfortable, (2) quite comfortable, (3) neutral, (4) uncomfortable, (5) very uncomfortable). Emotional reaction towards models has been specified as a *comfort* as our goal was to identify stimuli which are troubling and far from being enjoyable (see "Limitations of the Study" section for discussion of limitations of such a choice).

Overall, 237 participants took part in these four studies (137 in the laboratory studies and 100 in the online study). We present the summary of these study groups in Table 1.

**Table 1.** Study groups summary.

| Study | N | Age (mean) | Women | Form |
|---|---|---|---|---|
| Łupkowski et al. (2019) | 64 | 20.70±1.81 | 44 | laboratory |
| Ratajczyk et al. (2019) | 43 | 24.21±3.80 | 21 | laboratory |
| Łupkowski, Kuliński (2020) | 30 | 21.67±2.18 | 15 | laboratory |
| Łupkowski, Gierszewska (2019) | 100 | 27.78±8.28 | 55 | online |

In what follows, we use the aggregated data to answer some interesting questions related to the use of the set we used. As indicated in the introduction, these questions will be:

1. Do the online and laboratory studies have different results when it comes to human-likeness evaluation and emotional reaction?
2. Do we observe gender differences in the models' assessment?
3. Is there a correlation between the evaluation of human-likeness and time needed for the evaluation?

Answers to the above questions identify important factors related to the stimuli and are useful in formulating guidelines for their future use.

## Studies Versus Laboratory Studies

Firstly, we will analyse whether we observe differences in the evaluation of human-likeness and comfort of the models between online and laboratory conditions. We find such knowledge to be very important for future studies. In many cases, it is much easier to organise and conduct online studies. However, the price is the lack of control over the exact way of displaying the visual stimuli (i.e. different screen sizes, colour profiles, web-browsers, etc.) and other disturbing factors (e.g., a subject is disturbed by incoming notifications from another web-browser tab).

### Human-likeness

We use the human-likeness evaluation results to establish the alignment on the X axis of the UVH graph (see discussion in Łupkowski et al. 2017). As the measure of human-likeness score for a given robot we use the median of subjects evaluation. This measure was

successfully used in previous studies. Let us remind that the subjects answered the following question for each model: *How much does the presented model resemble a human?* (Answer on a scale 1-5, where: (1) Completely not human-like; (2) Rather not human-like; (3) Beginning to look human-like; (4) Rather human-like; (5) Completely human-like).

When we compare the online and the laboratory results, we can notice that they are significantly different for almost every model (see Table 2). However, when we consider only the median results for the robots, we may notice that they do not differ between laboratory and online conditions. This is illustrated in Figure 1. This result indicates that when we use median results for ordering the models, we can also decide to use the online form of study (and the form of presenting the models should not have an impact on the ordering).

**Table 2.** Online versus laboratory human-likeness evaluation. Mann-Whitney U test results (with Benjamini-Hochberg correction for multiple comparisons).

| Model | Test Results | |
|---|---|---|
| Red robot | U = 5864.5 | p = 0.006 |
| Woman | U = 5857.5 | p = 0.017 |
| Blue robot | U = 5864.5 | p = 0.026 |
| Battle robot | U = 5226.0 | p = 0.002 |
| Animated boy | U = 5343.0 | p = 0.002 |
| Silver robot | U = 5889.0 | p = 0.027 |
| Man in a cap | U = 5516.0 | p = 0.001 |
| Black-skinned man | U = 5395.5 | p = 0.001 |
| Animated girl | U = 5478.5 | p = 0.005 |
| Android woman | U = 5358.0 | p = 0.002 |
| Robot smiley | U = 4998.0 | p = 0.001 |
| Animated woman | U = 6306.0 | p = 0.129 |

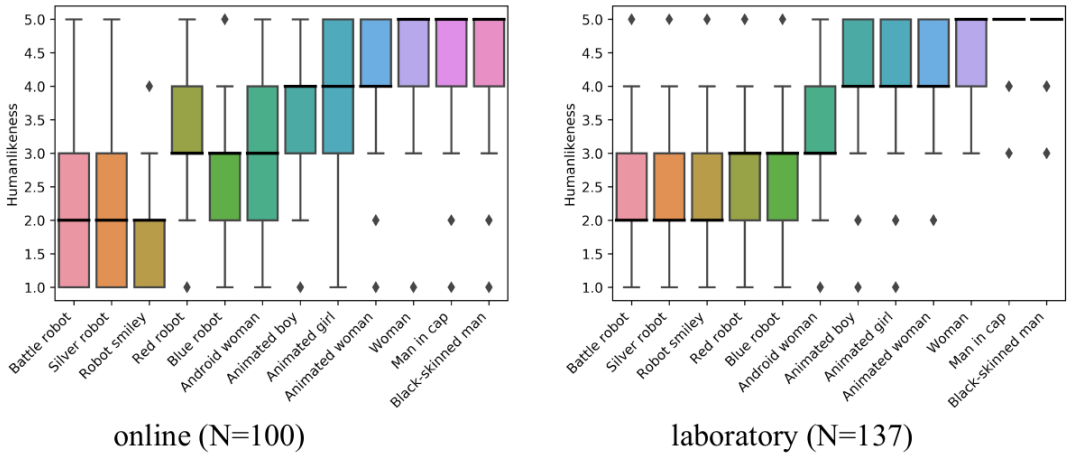


**Figure 1.** Online versus laboratory human-likeness evaluation (median results for robots).

This conclusion is supported by another observation. One may notice that the set can be divided into four sub-groups of models. We will refer to them as: (1) robots; (2) androids; (3) humanoids; and (4) humans. What is important is that the same sub-groups are visible across all four analysed studies (see Figure 2). This result is useful for future studies with the presented stimuli set, as we may potentially reduce the number of stimuli by selecting one representative stimulus for each established category, or, if needed, exclude a given category from the study (e.g., include only robots and androids as stimuli).

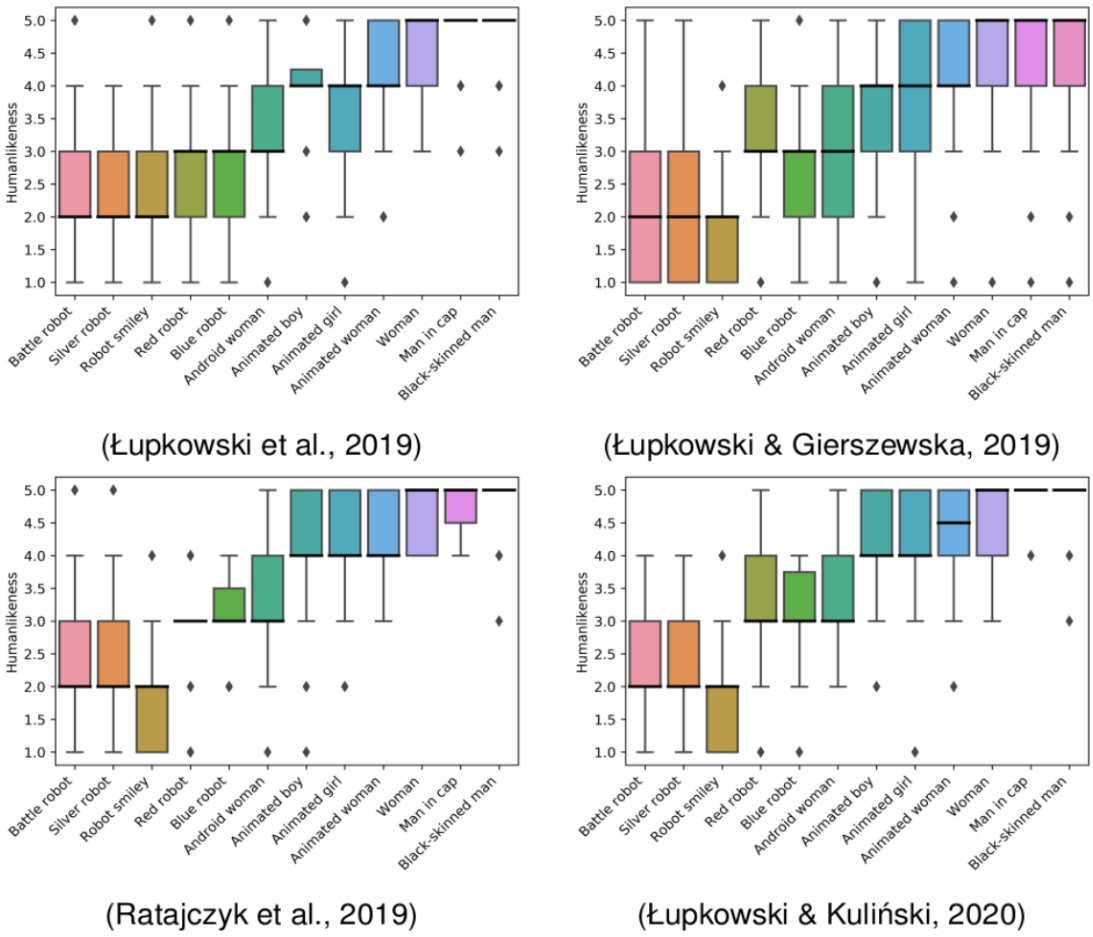


**Figure 2.** Sub-groups of the stimuli across the four studies.

## Comfort

For the comfort dimension, we compare the results of Łupkowski et al., 2019) and (Ratajczyk et al., 2019) (laboratory condition; N=107) with (Łupkowski, Gierszewska, 2019) (online condition; N=100).

As pointed in section "Studies selected for the analysis" subjects reported their comfort of observing each model by answering the following question: *How comfortable are you when you watch this model in a given environment?* (Answer on a scale 1-5, where: (1) very comfortable, (2) quite comfortable, (3) neutral, (4) uncomfortable, (5) very uncomfortable).

Table 3 presents differences between the comfort evaluation, which are not significant for most of the models, Two notable exceptions are the red robot and the blue robot. As in the above case of human-likeness, this result indicates that we may use the proposed set in both online conditions and in laboratory ones. The variable of comfort should not differ between these two conditions. See "Limitations of the Study" section for a broader discussion.

**Table 3.** Online versus laboratory comfort evaluation. Mann-Whitney U test results (with Benjamini-Hochberg correction for multiple comparisons).

| Model | Test Results | |
|---|---|---|
| Red robot | U = 3835 | p = 0.001 |
| Woman | U = 5184 | p = 0.377 |
| Blue robot | U = 4159 | p = 0.012 |
| Battle robot | U = 5078 | p = 0.340 |
| Animated boy | U = 5283 | p = 0.438 |
| Silver robot | U = 4934 | p = 0.269 |
| Man in a cap | U = 4980 | p = 0.283 |
| Black-skinned man | U = 4606 | p = 0.090 |
| Animated girl | U = 5120 | p = 0.348 |
| Android woman | U = 4891 | p = 0.267 |
| Robot smiley | U = 4583 | p = 0.090 |
| Animated woman | U = 4535 | p = 0.090 |

## Are There Any Gender Differences in the Assessment of the Models?

Based on the results reported in the HRI literature (see e.g, Pochwatko et al. 2015; Picarra, Giger, Pochwatko, and Goncalves 2016; Picarra, Giger, Pochwatko, and Mozaryn 2016) we expect to observe gender differences in the evaluation of our models. Picarra, Giger, Pochwatko, and Goncalves (2016) (see also Picarra, Giger, Pochwatko, and Mozaryn 2016) suggest that the possible explanation for these differences is that male and female participants associate robots with different contexts, such as the participant could potentially interact with robots in different settings, i.e. industrial vs. domestic robots or help with unemployment vs. help at home.

What is interesting in the aggregated data from our studies is that the differences between men and women are not significant for the human-likeness dimension evaluation. The only exception to this observation is for the model of the battle robot. As illustrated in Figure 3, men perceive it as more human-like than women.

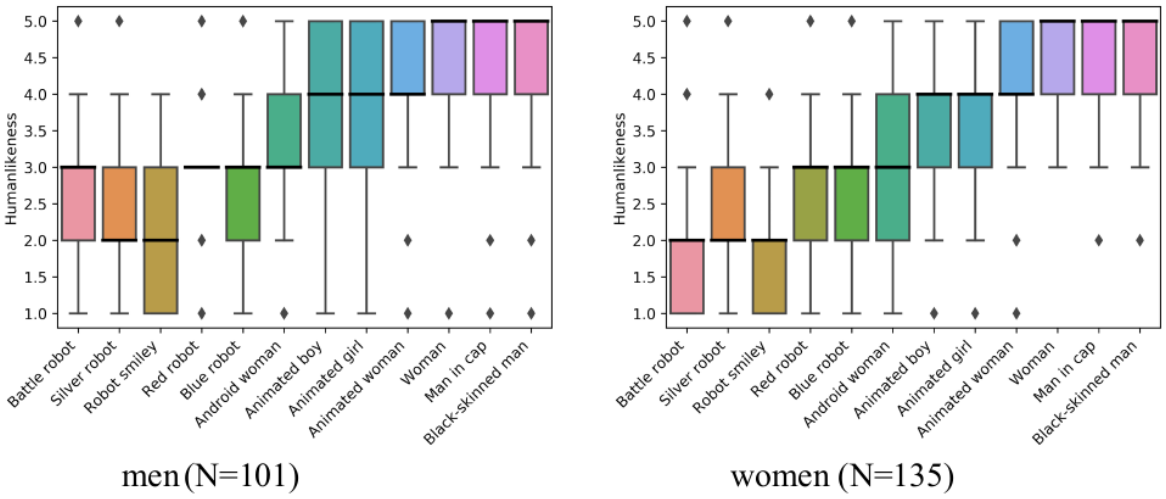


**Figure 3.** Men versus women in human-likeness evaluation.

As for the comfort evaluation, we do not observe any differences between men and women. Detailed test results are presented in Table 4.

**Table 4.** Men vs. women in human-likeness and comfort evaluation. Mann-Whitney U test results (with Benjamini-Hochberg correction for multiple comparisons).

| Model | H-L test results | Comf. test results |
|---|---|---|
| Battle robot | U = 3346, p < 0.001 | U = 5008, p = 0.498 |
| Red robot | U = 4712, p = 0.312 | U = 5043, p = 0.498 |
| Blue robot | U = 4225, p = 0.066 | U = 4672, p = 0.498 |
| Silver robot | U = 4469, p = 0.183 | U = 4896, p = 0.498 |
| Android woman | U = 4496, p = 0.183 | U = 4636, p = 0.498 |
| Robot smiley | U = 4604, p = 0.247 | U = 5027, p = 0.498 |
| Animated boy | U = 4803, p = 0.387 | U = 4571, p = 0.498 |
| Animated girl | U = 4902, p = 0.393 | U = 4908, p = 0.498 |
| Animated woman | U = 4088, p = 0.034 | U = 4708, p = 0.498 |
| Woman | U = 4948, p = 0.393 | U = 4990, p = 0.498 |
| Man in cap | U = 4969, p = 0.393 | U = 4871, p = 0.498 |
| Black-skinned man | U = 4879, p = 0.387 | U = 5054, p = 0.498 |

## Is There a Correlation Between the Human-likeness Evaluation and Time Needed for this Evaluation?

In two of the analysed studies (Ratajczyk et al., 2019) and (Łupkowski, Kulinski, 2020), the procedure included the measurement of the time taken in providing answers to questionnaire questions (N=73, laboratory condition). This measurement allows us to test whether we would observe a correlation between the time taken providing the answer to the question about human-likeness and the evaluation provided in the answer.

As illustrated in Figure 4, we observe a significant correlation between the mean time of human-likeness evaluation (in seconds) and median values of this evaluation – rS=-0.604, p=0.037. The less human-like a model was perceived, the more time a subject needed to provide his/her evaluation.

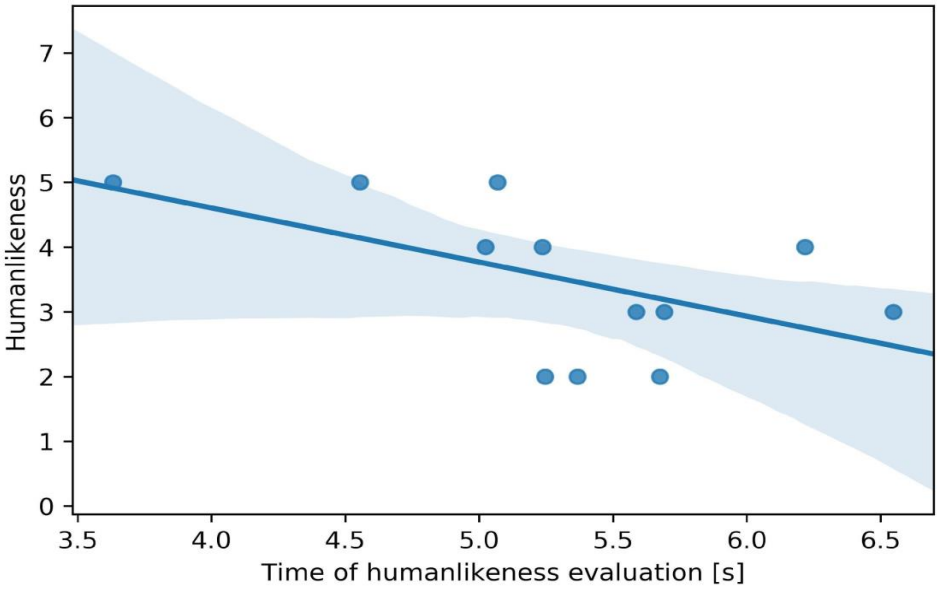


**Figure 4.** Correlation of human-likeness scores with time taken for the evaluation.

When it comes to the implications for the future use of the set, these results indicate that it is valuable to collect response times along with answers. These measurements will reveal interesting relationships between self-reported data and more objective measures.

We may also hypothesise that such a correlation suggests that the human-likeness scale is not uniform (see also the discussion in Katsyri et al. 2015 or Łupkowski et al. 2017). This means that the more leftmost a model on this scale, the more difficult the evaluation is. This difficulty is based on the model's appearance, as the least human-like models do not have visible facial features, or even hands or feet. This observation implies the usefulness of supplementing self-reported human-likeness evaluation with additional questions addressing aspects of the models which are important to the subjects while they are asked for human-likeness evaluation. Additional data (probably of qualitative characteristics) would provide us with a more in-depth understanding of the human-likeness evaluation process.

## ABOT DATABASE ROBOT HUMAN-LIKENESS PREDICTOR STUDY

In this section we present the evaluation of the set with the use of the ABOT (Anthropomorphic roBOT) Database (http://abotdatabase.info/). It is a collection of real-world anthropomorphic robots that has been created for both research and commercial purposes (Phillips, Zhao, Ullman, & Malle, 2018). Each robot in the database is described with the Human-likeness score and Dimension scores for: Body-Manipulators (torso, legs, arms, hands, fingers), Facial Features (head, face, eyes, mouth), and Surface Look (gender,

head hair, skin, nose, eyebrow, eyelashes, apparel). Exemplary robots along with their ABOT scores are presented in Figure 5. These factors were established as the result of the study described in (Phillips et al., 2018, p. 107-110). At the time of designing the set of models, we were not aware of the ABOT Database. However (as we have described in section "The Set") we took into consideration the following factors while choosing the models: facial features, presence of hands and feet, etc. Thus, for the needs of this paper, we have also decided to supplement our own choices and reported study results with the evaluation of the set against a very interesting ABOT tool designed for human-likeness prediction.

The ABOT Robot Human-Likeness Predictor (http://abotdatabase.info/predictor) is an online tool associated with this database. This tool allows the estimation of how human-like a given model will be perceived. After answering a series of questions for a given model (covering parameters like elements of clothing, skin appearance, the presence of hair, limbs description, etc.), the predictor provides the *Human-likeness score* and *Dimension scores.* With these scores, one can compare the model to the robots already present in the core collection of the ABOT Database[1]. Thus, we have decided to use the Predictor to estimate the human-likeness of our stimuli.

The detailed results are presented in Table 5. One may notice that according to the estimations, animated characters and human groups were evaluated as almost completely human-like (received H-L score is 94.55). This is understandable, as the model underlying these predictions was developed for commercial robots. As for robots and androids, we observe that their results vary and therefore provide more insights for these sub-groups of models. For the robots group, we observe that the lowest estimated H-L score is for the Silver-robot (all the dimensions are also the lowest for this model, as it has no hands and feet, no clothes or skin, and its facial features are only suggested). Then, we have the Robot Smiley with 10 points more, which is in turn followed by the battle robot, which scores more points. For the android group, we do not observe the same differences in the H-L score, as they received over 60 points. These models do not differ in the body manipulators' dimension and have a similar surface appearance. When it comes to facial features, only the Android Woman, who has a human-like facial feature, is found.


**Figure 5.** Exemplary ABOT Database entries.

**Table 5.** Detailed results of the ABOT Predictor for the set of models.

| Model | Body manipulators | Surface look | Facial features | H-L Score |
|---|---|---|---|---|
| Silver robot | 0.60 | 0.00 | 0.50 | 34.18 |
| Robot smiley | 1.00 | 0.00 | 0.50 | 44.22 |
| Battle robot | 1.00 | 0.14 | 0.75 | 57.59 |
| Blue robot | 1.00 | 0.14 | 0.50 | 62.39 |
| Red robot | 1.00 | 0.15 | 0.50 | 62.52 |
| Android woman | 1.00 | 0.29 | 1.00 | 65.44 |
| Animated boy | 1.00 | 1.00 | 1.00 | 94.55 |
| Animated girl | 1.00 | 1.00 | 1.00 | 94.55 |
| Animated woman | 1.00 | 1.00 | 1.00 | 94.55 |
| Woman | 1.00 | 1.00 | 1.00 | 94.55 |
| Man in cap | 1.00 | 1.00 | 1.00 | 94.55 |
| Black-skinned man | 1.00 | 1.00 | 1.00 | 94.55 |

The ABOT H-L score allows for establishing only three sub-groups for the stimuli set – humanoids and humans (described in subsection "Human-likeness") are merged into one group with the H-L score of 94.55. This may be explained in the light of the main purpose of the ABOT Database Predictor – it was designed on the basis of existing robots. Sub-groups of robots (Silver Robot, Robot Smiley, Battle Robot) and androids (Blue Robot, Red Robot, Android Woman) are preserved. Notably, exactly the same models are included in these sub-groups as the ones presented in Figure 2.

## SELECTED STUDIES AND THEIR IMPLICATIONS FOR THE STIMULI USAGE

### Context Condition

The main inspiration for studying the context condition comes from Hanson's (2005) critique of the uncanny valley phenomenon. The claim is that a well designed humanoid robot approached in a relevant environment will not trigger a negative reaction. Also, the intuitive approach to the uncanny valley hypothesis, especially for the case of computer generated models, suggests that models would appear more familiar and acceptable when presented in an expected and known context (i.e., from science fiction animations or computer games).

The context condition of the set was studied in (Łupkowski et al., 2019) and (Ratajczyk et al., 2019). For these experiments, two sets of stimuli were prepared: (A) models on a neutral background (an empty room with uniformly coloured greyish walls); and (B) models on a background that was suitable for the given model. By suitable we mean a background presenting a context that is relevant to a given model. Robots were placed in the industrial context of a factory, androids in science-fiction scenery, animated characters on a cartoon-like background, and humans in a background depicting a city. The goal was to place the model in a background that the model would not be surprising to see. Examples are presented in Figure 6.

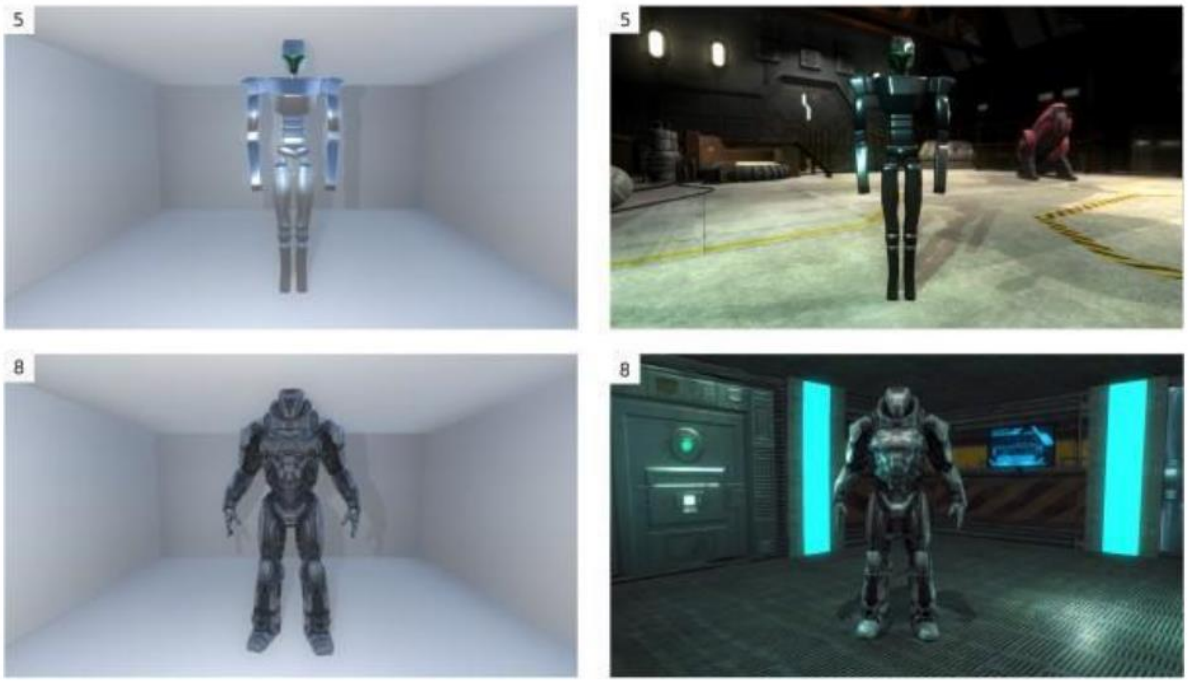


**Figure 6.** Exemplary stimuli used for the context condition study.

For (Łupkowski et al., 2019), we have checked the differences in (declared) comfort scores and (declared) emotional reactions between the two sets of stimuli. In (Ratajczyk et al., 2019), the difference was measured in terms of electrodermal activity (EDA). Based on these results, both studies ended with the conclusion that we observe no statistical differences between models from groups (A) and (B). No influence of the background on the perception of models was observed.

A practical application of this result is that there is no need to render complex images for future studies using our stimuli. As our study shows, models may be used on a uniform background (of an empty room) and are already available in this form. In our opinion, this solution is also more convincing, as there are no elements in the background which may disturb a subject. Naturally, we agree that further studies concerning the background condition for the uncanny valley phenomenon are needed – especially with real robots (see e.g. von der Pütten, Krämer, Becker-Asano, & Ishiguro, 2011; Zlotowski et al., 2015).

## Perceived Characteristics of the Models

In the study (Łupkowski, Gierszewska, 2019), we asked our subjects questions, which were addressing the perceived characteristics of the models. These questions were of the following form: *Mark on a given scale to which extent you would agree that the presented model is trustworthy/hostile/strong*. Answers were provided on a 5-point scale: 1—totally disagree to 5 —totally agree. Detailed results are presented in Table 6.

**Table 6.** A comparison of models' features (medians; Łupkowski, Gierszewska 2019)

| Model | Intelligent | Trustworthy | Hostile | Strong |
|---|---|---|---|---|
| Red robot | 3.0 | 3.0 | 2.0 | 3.0 |
| Woman | 3.0 | 3.0 | 2.0 | 3.0 |
| Blue robot | 3.0 | 3.0 | 3.0 | 4.0 |
| Battle robot | 3.0 | 2.0 | 1.5 | 2.0 |
| Animated boy | 3.0 | 3.0 | 1.5 | 2.0 |
| Silver robot | 3.0 | 3.0 | 3.0 | 3.0 |
| Man in cap | 3.0 | 3.0 | 3.0 | 3.0 |
| Black-skinned man | 3.0 | 3.0 | 3.0 | 3.0 |
| Animated girl | 3.0 | 3.0 | 2.0 | 2.0 |
| Android woman | 4.0 | 3.0 | 3.0 | 4.0 |
| Robot smiley | 2.0 | 2.0 | 3.0 | 4.0 |
| Animated woman | 3.0 | 3.0 | 3.0 | 3.0 |

We interpret answers where median is equal 3 as showing that a given feature is not relevant for the model. For the intelligence feature, the Android woman received the highest scores. On the other hand, we observe the lowest scores for the Robot smiley. As for the trustworthiness of the models, we observe lowered scores for the Battle robot and Robot smiley. When it comes to the hostility feature, we observe that the Red robot, Woman, Man in a cap, and Animated girl were all perceived as rather not hostile. The Animated boy received the lowest scores (median=1.5), which is in line with the results for the emotional reaction presented in Table 7. For the strength feature models Robot smiley, Android woman, and Blue robot were scored higher than others. The exceptional model here is the Battle robot, which received a maximal score for this feature. This observation is also in line with the reported emotional reaction for that model.

**Table 7.** The summary of dominant declarative emotional reactions for models (Łupkowski & Gierszewska, 2019).

| Model | The dominant reaction | % of declarations |
|---|---|---|
| Red robot | neutral | 65% |
| Woman | neutral | 45% |
| Blue robot | neutral | 57% |
| Battle robot | neutral | 76% |
| Animated boy | friendly | 63% |
| Silver robot | neutral | 39% |
| Man in cap | neutral | 55% |
| Black-skinned man | neutral | 50% |
| Animated girl | strange | 39% |
| Android woman | anxiety | 39% |
| Robot smiley | anxiety | 44% |
| Animated woman | neutral | 51% |

As mentioned above, in our study we also asked about a declared emotional reaction for the models. Subjects answered the question: *What is your emotional reaction for the presented model?*. Subjects could choose maximally two among four possible answers: (i) "the model seems to be friendly"; (ii) "my reaction is neutral"; (iii) "the model looks strange"; (iv) "the model makes me feel anxious". The summary of the dominant declarative emotional reactions for the models is presented in Table 7.

For the majority of models, the reaction is rated as neutral. The Animated boy is clearly perceived as friendly. Moreover, it is the only model with this evaluation result. This result is in line with perceived features: a low level of hostility and a low level of strength. The Animated girl is rated as strange, which, as we discuss in (Łupkowski, Gierszewska, 2019, p. 110), may be the result of an imperfect rendering of the model (it is worth to stress that the model was corrected in the re-rendering process for the set published along with this paper). As for the anxiety response, this was a dominant declared emotional reaction for the Robot smiley, Android woman, and especially for the Battle robot (rated also as the strongest model, highly hostile, and with a low trustworthiness level).

## LIMITATIONS OF THE STUDY

We are aware of certain limitations of the presented study – which are mainly a consequence of the idea of analysing the aggregated data from previous studies and certain limitations of them. First of all, as an indicator of emotional reaction towards stimuli, we use only the *comfort* scale. The main reason for such a (limited) choice was to utilise an analysis of results from completed and published studies that used the same set of stimuli, as well as the same questions for participants. In the future, it would be valuable to examine the stimuli with other UVH-focused scales, such as *eeriness* and *likability*.

Secondly, the set of models does not contain fully photo-realistic humanlike characters. The research indicated that highly photo-realistic humanlike models are required to examine the full shape of the uncanny valley (Katsyri et al., 2015; Palomaki et al., 2018). However, semi-realistic models seem to be sufficient to reveal the decrease in likability (Bartneck, Kanda, Ishiguro, & Hagita, 2007; Burleigh, Schoenherr, & Lacroix, 2013; Katsyri et al., 2017). In the current paper, we focused on the examination of differences between online and laboratory studies, gender differences in models assessment and correlation between human-likeness evaluation and time needed for this.

## SUMMARY

In this paper, we have presented a set of stimuli for several UVH-related studies. We describe and discuss the results of analyses performed on the aggregated data from previous studies in which this set has been used. We also used some of the particular results reported in these studies. Additionally, we used the ABOT Database Prediction tool to evaluate human-likeness scores for the models in the set. On the basis of our results, we formulate the following guidelines for the future use of this set:

## A. Models

1. The set forms four sub-groups on the human-likeness dimension: robots; androids; humanoids; humans. Within these sub-groups, median scores for the models are all equal.
2. It is possible to select a representative model for each sub-group. One may focus on the models with certain interesting characteristics (as summarised in Table 6) or emotional evaluation (Table 7).
3. It is also possible to omit a given sub-group of models, e.g. excluding the humanoids sub-group to focus only on the robot-android-human continuum. This is also suggested by the ABOT Predictor analysis (as humanoids reach human scores in this tool).
4. ABOT Predictor scores allow for the use of our models in addition to the ABOT Database robots for future UVH-related studies (as we have shown how they relate to each other on the human-likeness continuum).

## B. Stimuli presentation

1. Models need not be embedded within an advanced background. They may be presented in an empty room to avoid potential distractions.
2. The set may be used successfully within both laboratory and online study scenarios.
3. Models should be presented one by one, preferably in a randomised order.
4. The negative correlation that was observed for the human-likeness evaluation and the time needed to answer questions suggests that the time measure is a valuable additional measure for future studies. What is more, it would be beneficial to use additional questions, which address factors that are taken into account by subjects while evaluating human-likeness.

## C. Study groups

1. We should bear in mind that the aggregated data used in this paper comes from a relatively narrow age group (see Table 1). Results may vary for other age groups, and certainly further studies addressing age variation are needed.
2. We have not observed any gender differences for the human-likeness evaluation for the majority of the models, and no gender differences for the comfort evaluation. This result needs further studies, but also suggests that the studies with the set need not otherwise be balanced for gender.

As we stated in the Introduction, we believe that the publicly available pre-validated set of stimuli described here along with the guidelines formulated above enhances future studies aimed at replicating the uncanny valley effect (for computer generated graphics).

## ENDNOTES [NOT FOOTNOTES]

1. The validation study for the ABOT Predictor is presented in (Phillips et al., 2018, p. 110-111). It is also worth mentioning that the ABOT Database entries were used as the basis for the evaluation of UVH (see Kim, Bruce, Brown, de Visser, & Phillips, 2020).

## REFERENCES


Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2007). Is the uncanny valley an uncanny cliff?. In *Ro-man 2007-the 16th ieee international symposium on robot and human interactive communication* (pp. 368-373).

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, 29(3), 759-771.

Dill, V., Flach, L. M., Hocevar, R., Lykawka, C., Musse, S. R., & Pinho, M. S. (2012). Evaluation of the uncanny valley in CG characters. In *Intelligent virtual agents* (pp. 511-513).

Geller, T. (2008). Overcoming the uncanny valley. *IEEE computer graphics and applications*, 28(4).

Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., ... Stephanou, H. (2005). Upending the uncanny valley. In *Proceedings of the 20th national conference on artificial intelligence* - volume 4 (pp. 1728-1729). AAAI Press. https://doi.org/10.5555/1619566.1619636.

Katsyri, J., Makarainen, M., & Takala, T. (2017). Testing the 'uncanny valley' hypothesis in semi realistic computer-animated film characters: An empirical evaluation of natural film stimuli. *International Journal of Human-Computer Studies*, 97, 149-161.

Kim, B., Bruce, M., Brown, L., de Visser, E., & Phillips, E. (2020). A comprehensive approach to validating the uncanny valley using the anthropomorphic robot (ABOT) database. In *2020 systems and information engineering design symposium* (SIEDS) (pp. 1-6).

Katsyri, J., Forger, K., Makarainen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology, 6*, 390. https://doi.org/10.3389/fpsyg.2015.00390.

Lupkowski, P., & Gierszewska, M. (2019). Attitude towards humanoid robots and the uncanny valley hypothesis. *Foundations of Computing and Decision Sciences, 44*(1), 101-119.

Łupkowski, P., & Kulinski, C. (2020). Aspekty percepcji wirtualnych modeli humanoidalnych w kontekście zjawiska doliny niesamowitości. [Perceiving human-like humanoid models—case study for the uncanny valley studies] (In preparation).

Łupkowski, P., Rybka, M., Dziedzic, D., & Włodarczyk, W. (2017). Human-likeness assessment for the uncanny valley hypothesis. *Bio-Algorithms and Med-Systems, 13*(3), 125-131. https://doi.org/10.1515/bams-2017-0008.

Łupkowski, P., Rybka, M., Dziedzic, D., & Włodarczyk, W. (2019). The background context condition for the uncanny valley hypothesis. *International Journal of Social Robotics, 11*(1), 25-33. https://doi.org/10.1007/s12369-018-0490-7.

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? uncanny responses to computer generated faces. Computers in human behavior, 25(3), 695-710.

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition, 146*, 22-32.

Mathur, M. B., Reichling, D. B., Lunardini, F., Geminiani, A., Antonietti, A., Ruijten, P. A., Levitan, C.A., Nave, G., Manfredi, D., Bessette-Symons, B., Szuts, A., Aczel, B. (2020). Uncanny but not confusing: Multisite study of perceptual category confusion in the uncanny valley. *Computers in Human Behavior, 103*, 21-30. https://doi.org/10.1016/j.chb.2019.08.029.

McDonnell, R., Breidt, M., & Bülthoff, H. H. (2012). Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG), 31*(4), 1-11.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98-100. (Original work published in 1970 in Japanese).

Palomaki, 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. https://doi.org/10.1016/j.heliyon.2018.e00939.

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (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).

Piçarra, N., Giger, J.-C., Pochwatko, G., & Gonçalves, G. (2016). Making sense of social robots: A structural analysis of the layperson's social representation of robots. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 66*(6), 277-289.

Piçarra, N., Giger, J.-C., Pochwatko, G., & Mozaryn, J. (2016). Designing social robots for interaction at work: socio-cognitive factors underlying intention to work with social robots. *Journal of Automation Mobile Robotics and Intelligent Systems, 10*(4), 17-26.

Piwek, L., McKay, L. S., & Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition, 130*(3), 271-277.

Pochwatko, G., Giger, J.-C., Rozanska-Walczuk, M., Swidrak, J., Kukielka, K., Mozaryn, J., & Pigarra, N. (2015). Polish version of the negative attitude toward robots scale (nars-pl). *Journal of Automation Mobile Robotics and Intelligent Systems, 9*(3), 65-72.

Ratajczyk, D., Jukiewicz, M., & Lupkowski, P. (2019). Evaluation of the uncanny valley hypothesis based on declared emotional response and psychophysiological reaction. *Bio-Algorithms and Med-Systems, 15*(2). https://doi.org/10.1515/bams-2019-0008.

Ratajczyk, D. (2022). Shape of the Uncanny Valley and Emotional Attitudes Toward Robots Assessed by an Analysis of YouTube Comments. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00905-x.

Schneider, E., Wang, Y., & Yang, S. (2007). Exploring the uncanny valley with Japanese video game characters. In: *Proceedings of DiGRA International Conference: Situated Play*. Tokyo, Japan, September 2007. DIGRA, pp. 546-549.

Schwind, V., Wolf, K., & Henze, N. (2018). Avoiding the uncanny valley in virtual character design. *Interactions, 25*(5), 45-49.

von der Pütten, A. M., Krämer, N. C., Becker-Asano, C., & Ishiguro, H. (2011). An android in the field. In Proceedings of the 6th international conference on human-robot interaction (pp. 283-284).

Zhang, J., Li, S., Zhang, J.-Y., Du, F., Qi, Y, & Liu, X. (2020). A literature review of the research on the uncanny valley. In *International conference on human-computer interaction* (pp. 255-268).

Zibrek, K., & McDonnell, R. (2014). Does render style affect perception of personality in virtual humans?. In *Proceedings of the ACM symposium on applied perception* (pp. 111-115).

Zlotowski, J. A., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2015). Persistence of the uncanny valley: The influence of repeated interactions and a robot's attitude on its perception. *Frontiers in psychology, 6*, 883. https://doi.org/10.3389/fpsyg.2015.00883.

## Authors' Note

Work on this project was founded within the Faculty of Psychology and Cognitive Science AMU 2020 grant.

All correspondence should be addressed to
Paweł Łupkowski
Faculty of Psychology and Cognitive Science
Adam Mickiewicz University
Szamarzewskiego 89a
60-568 Poznań, Poland
Pawel.Lupkowski@amu.edu.pl

Dawid Ratajczyk
Faculty of Psychology and Cognitive Science
Adam Mickiewicz University
Szamarzewskiego 89a
60-568 Poznań, Poland
Dawid.Ratajczyk@amu.edu.pl

*Human Technology*
ISSN 1795-6889
https://ht.csr-pub.eu