Int. J. Human–Computer Studies 196 (2025) 103443

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Item-level implicit affective measures reveal the uncanny valley of robot faces☆,☆☆

Motonori Yamaguchi

Department of Psychology, University of Essex, Colchester, United Kingdom

ARTICLE INFO

Keywords:
Uncanny valley
Artificial intelligence
Affective priming
Single-category IAT

ABSTRACT

As the opportunity to interact with humanoid robots and virtual avatars increases, the emotional impact of the interaction with these artificial agents becomes an important consideration. The uncanny valley effect is a psychological phenomenon relevant to such a consideration. Although the uncanny valley remained untested for several decades, recent empirical studies confirmed the uncanny valley effect when human observers rated their liking of robots' faces. To uncover the uncanny valley in behavioral measures of affective response, the present study used two implicit affective tasks, affective priming and single-category implicit association test (IAT). Positivity scores for each of the images of robot faces were derived and were plotted against the humanness rating of the robot faces. The results demonstrated the uncanny valley effect in these implicit behavioral measures. The finding indicates the effectiveness of using these implicit measures to assess affective responses to individual items rather than to groups of items, and it suggests the potential of these behavioral paradigms for wider application outside laboratory research.

1. Introduction

In the last decade, there have been drastic changes in the landscape surrounding AI technologies, making it feasible to develop artificial agents such as humanoid robots and digital avatars that exhibit human-like communication and human-level performance in complex tasks. Many service providers now use a chatbot as the initial point of online customer contact, and more home appliances embed a virtual assistant technology that communicates with users in a human-like manner. Public and private sectors have also started using robots or robot-like consoles for front-facing services to communicate with their users and customers. This recent explosion of advanced AI and robotics technologies offers a valuable opportunity to scrutinize their psychological impact on users, especially in the socio-emotional domain. One way to assess such impact is to use behavioral paradigms that measure people's implicit affective responses.

The method of measuring implicit affective responses, or implicit attitudes, became a subject of extensive research in the 1980s, which resulted in several different experimental paradigms thereafter (e.g., Bar-Anan and Nosek, 2014; De Houwer, Teige-Mocigemba, Spruyt, and Moors, 2009). Of these, the two most popular techniques have been the *affective priming* procedure (Fazio, Sanbonmatsu, Powell, and Kardes, 1986) and the *implicit association test* (IAT; Greenwald, McGhee, and Schwartz, 1998), both of which have been in extensive use in psychological research. The present study examined the effectiveness of these behavioral paradigms to measure people's affective reactions to artificial agents (humanoid robots). To do so, I tested whether these measures could reveal a widely acknowledged, but poorly understood, phenomenon in human-robot interaction, known as the *uncanny valley effect* (Mori, 1970).

A challenge of using these behavioral paradigms is that an evaluation of the uncanny valley effect entails measuring affective reactions at an individual item level, as opposed to those at a category level (i.e., 'robot' vs. 'human') for which these paradigms are designed and have been used. The present study reports one replication, which evaluated the uncanny valley effect in a self-rating likability scale of robot faces, and two new behavioral experiments, which used the affective priming procedure and a version of the IAT, called the single-category IAT

☆ For the purpose of open access, the author has applied a Creative Commons Attribution (CC-BY) licence to any Author Accepted Manuscript version arising. The data, analysis scripts, and stimuli used in the present study are available for reanalysis purposes from the Open Science Framework project page (https://osf.io/3ae2k/).

☆☆ The author thanks Timofei Liukshin for his assistance in the preparation of the materials and setting up the online survey for Experiment 1, Ethan Richards, Fabien Coomber, and I-Cheng Liu, for their assistance in data collection for Experiments 2 and 3. The project was supported partly by a seedcorn grant from ESSEXLab, University of Essex.

E-mail address: motonori.yamaguchi@essex.ac.uk.

https://doi.org/10.1016/j.ijhcs.2024.103443
Received 3 November 2023; Received in revised form 18 December 2024; Accepted 31 December 2024
Available online 2 January 2025

(Karpinski and Steinman, 2006), to reveal the uncanny valley effect in affective reactions to individual robot faces.

1.1. The uncanny valley effect

The uncanny valley effect is an observation that people's liking of robots or digital avatars (such as computer graphic characters in animated films) has a unique non-linear relationship with the level of human-likeness of the robots and avatars (see Fig. 1). This idea was first proposed by Mori (1970), who observed that people's affinity for a robot increases as its design approached the human appearance, but the affinity drops dramatically when the appearance of a robot became highly similar to a human. Mori called this drop of affinity for human-like robots the uncanny valley. He suggested that people experienced an unsettling feeling when they encountered such artifacts as prosthetic hands and puppets that were made to look similar to human bodies but did not achieve the perfection of the human appearance. For several decades, the uncanny valley was widely known and cited, but it remained to be untested until recently (for reviews, see Diel, and MacDorman, 2021; Kätsyri et al., 2015; Wang et al., 2015).

The original conception of the uncanny valley left many ambiguities as to how it could be operationalized for empirical tests. The horizontal axis of Mori's original illustration of the uncanny valley was represented by the degree of similarity to human (Human-likeness in Fig. 1), which has been manipulated or controlled in different manners, such as morphing images of a human and a non-human character (Hanson, 2006; Seyama and Nagayama, 2007), distorting facial features (MacDorman and Chattopadhyay, 2016), creating mismatch between facial expressions and voice (Mitchell et al., 2011; Tinwell et al., 2013), and using pictures or movies of a variety of existing robots (Mathur and Reichling, 2016) or animated characters (MacDorman, 2006; MacDorman and Entezari, 2015). Many previous studies used an arbitrary set, and often a limited range of, exemplars that might have failed to cover the continuum of human-likeness sufficient to observe the phenomenon (Lischetzke et al., 2017; Tinwell et al., 2011).

The vertical axis of Mori's illustration of the uncanny valley was represented by a Japanese term "*Shinwakan*" which have been translated into different English terms, such as "familiarity" (MacDorman and Ishiguro, 2006), "comfort" (K.F. MacDorman et al., 2009b), "warmth" (Mitchell et al., 2011), "affinity" (Mori, 1970/2012), "likability" (Ferrey et al., 2015), "valence" (Cheetham et al., 2011), and reverse-scaled "eeriness" (Ho & MacDorman, 2016). The idiosyncrasies of the measurements used in the previous studies of the uncanny valley effect are partly responsible for mixed evidence for (e.g., MacDorman and Ishiguro, 2006; Mathur et al., 2020) and against (e.g., Bartneck and Kanda, 2009; Cheetham et al., 2013) Mori's hypothesis. The lack of coherence of the measurement of affinity and of the manipulation of human-likeness underscores a need of further investigations on how best the uncanny valley effect could be tested empirically.

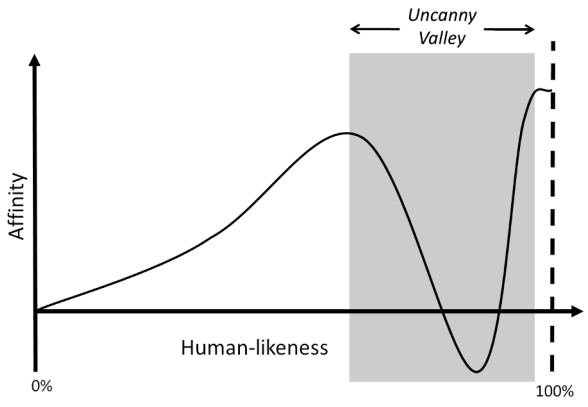

Fig. 1. An illustration of a hypothetical Uncanny Valley Effect (reproduced as depicted in Mori's original essay).

Diel et al. (2021) surveyed various methods used to investigate the uncanny valley effect and noted that artificially modified images to construct the human-likeness typically involves implicit assumptions about what underlies the perception of human-likeness. These morphed images would result from mixing a small number of images (typically two or three images, one robot face and one human face at the two extremities with or without a humanoid face in-between), selection of which can determine the appearance of morphed images. While this technique allows precise control of the proportions of human and non-human appearances, morphed images can also involve an issue of misalignment between morphed images, such that unusual facial features (e.g., having two noses) could be created (Yamada et al., 2013), causing negative affect independently of human-likeness. An alternative method is to use distinct entries of exemplars that involve a large number of specific instances of existing robot faces. A successful case of this type is Mathur and Reichling's (2016) study, in which the researchers used a sample of 80 pictures of the existing robots' faces and obtained a self-rating scale of human-likeness ranging from –100 (mechanical) to 100 (human). A self-rating scale of likability of the images (which also ranged from –100 to 100) was plotted against the human-likeness scale and demonstrated the uncanny valley effect. This stimulus set was also used in another study conducted by a different group of researchers, which replicated the uncanny valley effect with different age groups (Tu et al., 2020). The stimulus set was further expanded by the original authors in a subsequent many-lab study that also obtained the uncanny valley effect with a self-rating likability scale (Mathur et al., 2020). The present study adopted the expanded stimulus set of Mathur et al.'s to manipulate the human-likeness scale.

As in Mathur et al.'s (2020) study, most previous studies used self-reporting scales of affinity. Other indirect behavioral measures of affinity were also adopted in some studies. For example, Mathur et al. included a mouse-tracking measurement in a categorization task for which participants classified stimuli into the category *human* or *robot*. Other studies also used behavioral measures, such as eye-tracking (Cheetham et al., 2013; Minato et al., 2006), preferential looking (Matsuda et al., 2012), avoidance response (Strait et al., 2017), and motor interference (Kilner et al., 2003). However, the interpretations of these behavioral measures are often difficult because they could reflect factors other than affective reactions (e.g., perceptual difficulty). Arguably, more direct measures of affective reaction would include the aforementioned behavioral paradigms, such as affective priming and the IAT. K.F. MacDorman et al. (2009a) used the IAT to examine the attitudes toward robots among Japanese and US populations, but this study did not evaluate the uncanny valley effect. Instead, they tested whether weapons were associated more strongly with robots than with humans, which they took as reflecting the degree of threat of robots and that of humans. The authors also noted a problem of using 'human' as an alternative to 'robot' in the IAT or other self-report questionnaires and behavioral paradigms, because participants' attitudes toward humans could greatly vary across individuals or cultures. For example, some may consider humans to be threatening (e.g., to the global climate or ecosystems), which may result in a higher positivity score for robots in comparison. The dichotomy between human and robot involved in the categorization task used by Mathur et al. (2020) or in the standard format of the IAT might not provide an accurate measure of affinity for robots. This is a common problem in many behavioral paradigms of implicit attitudes (O'Shea and Wiers, 2020). An absolute affective measure that is free from competing categories is required to evaluate people's affective reaction to robots.

In the present study, I considered two behavioral paradigms that do not involve the dichotomy between two competing categories. The first paradigm is affective priming (Fazio et al., 1986). This involves presenting two stimuli in rapid succession, where the first stimulus serves as a prime and the second as the probe. The probe is an evaluative stimulus (such as positive and negative words), and participants' task is to

identify whether the probe is a positive or negative stimulus. The prime is an affect-inducing stimulus that does not require a response. The prime is presented briefly and influences how quickly and accurately participants could respond to the probe, depending on whether the induced affect is congruent with the required response to the probe: If participants react to the prime positively, responses to the positive probe are facilitated, while responses to the negative probe are interfered with; the opposite is true if participants react to the prime negatively. This procedure does not require competing categories, because affective reaction to the prime is measured as the differences in response time (RT) and accuracy to positive and negative probes following a particular prime. The differences in RT and accuracy are called the *affective priming effect*.

The second paradigm is the single-category IAT (Karpinski and Steinman, 2006; also see Bluemke and Friese, 2008). The standard format of the IAT (Greenwald et al., 1998) involved two different categorization tasks, one that requires categorizing positive and negative words (as in affective priming) and the other that requires another set of stimuli (target stimuli) with two opposing classes (e.g., *robot* vs. *human*). The IAT effect refers to the difference in RT and accuracy between two conditions for which two target classes share the response keys with the positive and negative words. For example, if participants perceive robots to be more negative than humans, RT would be shorter (and accuracy would be higher) when robots share the response key with negative words than when they share the response key with positive words. However, because the categorization involves discrimination between human and robot, the effect could be driven solely by the attitude toward humans rather than that toward robots. The single-category IAT was meant to overcome this problem of competing categories. In this version of the IAT, the target stimuli consist only of exemplars from one class (robot), and participants' task is to respond to robots by pressing the key assigned to positive words in one block or the key assigned to negative words in the other block. Hence, no discrimination between competing categories is involved in the single-category IAT.

Although the affective priming procedure and the single-category IAT overcome the problem of competing categories, posed by K.F. MacDorman et al. (2009a) and others (e.g., Karpinski and Steinman, 2006; O'Shea and Wiers, 2020), they still face another challenge for the present purpose, which is that these paradigms have only been used to evaluate affective reactions to a category of stimuli but never to evaluate affective reactions to individual items. This poses a challenge because the uncanny valley effect can only be evaluated by plotting affinity scores of individual robots against their human-likeness scores. Hence, a behavioral measure for individual items is required, but it is not clear whether the affective priming procedure and the single-category IAT are sufficiently sensitive to capture the variability in affective responses to individual items within a category. Therefore, the present study tested whether these two behavioral measures can reveal the uncanny valley effect for robot faces.

1.2. The present study

The present study conducted three experiments. Experiment 1 attempted to replicate the uncanny valley effect in self-rating scale of likability of robot faces used in Mathur et al.'s (2020) study, before Experiments 2 and 3 that used the two different behavioral paradigms to measure affective responses to the robot faces. The replication was conducted in an online questionnaire asking participants to rate each of the robots' images on two scales; (1) "how machine- or human-like the robot looks" (machine-human, or MH, scale) and (2) "how pleasant or creepy it would be if you were to interact with them in an everyday situation such as asking questions in a museum information booth" (likability scale). To evaluate the uncanny valley, I adopted the data analysis strategy used by Mathur and Reichling (2016) that were also adopted by Tu et al. (2020). This entailed calculating MH and likability scores for each of the robot faces and fitting polynomial regression models to predict the likability scores based on polynomial terms of the MH scores (i.e., MH, MH^2, MH^3,…). To find which of these polynomial models was most parsimonious to describe the relationship between the likability and MH scores, model fits were compared in terms of F-tests. The parsimonious model was defined as the lowest degree model that fit to the data significantly better than a lower degree model while a higher degree model did not improve the fit. This replication study was essential to confirm that the stimulus set did result in the uncanny valley effect if the behavioral paradigms were to be effective to reveal it.

To give a preview, the results of Experiment 1 successfully replicated the uncanny valley effect in the subjective likability rating. Fifty robot faces were selected from Experiment 1 and used in Experiments 2 and 3 to test whether the two behavioral paradigms could reveal the uncanny valley in affective responses (see Fig. 2 for selected robot faces, and the Method of Experiment 2 for the selection procedure). Experiment 2 used the affective priming procedure, and Experiment 3 used the single-category IAT. These paradigms are illustrated in Fig. 3. In the affective priming procedure, participants responded to words (e.g., *happy, agony*) by judging whether its meaning was positive or negative by pressing one or the other key on the keyboard. The word was preceded by a photograph of a robot face, which only appeared briefly and disappeared as a word appeared. The differences in RT and response accuracy (i.e., percentages of error trials, or PE) between positive and negative words represented the affective priming effect for that robot. The positivity score was calculated by subtracting RT and PE for positive words from RT and PE for negative words, which should be larger for robots that were perceived to be more positive. In the single-category IAT, participants were presented with either a word (word trial) or a photograph of a robot face (target trial). These trials were randomly intermixed. Participants responded to words by pressing one or the other key to indicate whether its meaning was positive or negative, whereas they responded to all robot faces by pressing the key assigned to positive words in certain blocks of trials and by pressing the key assigned to negative words in other blocks. The differences in RT and PE on target trials (i.e., responses to robot faces) between blocks requiring the positive keypress and blocks requiring the negative keypress represented the IAT effects for those robot faces. The positivity score was calculated by subtracting RT and PE for the blocks with positive keypress from RT and PE for the blocks with negative keypress, which again should be larger for robots that were perceived to be more positive. To examine the uncanny valley effect, the positivity score was derived for each of the robot faces and averaged across participants, which was submitted to the same analysis procedure as in Experiment 1, except that the likability score was replaced by the positivity scores from the respective experiments.

2. Experiment 1

2.1. Method

2.1.1. Participants

Two hundred and seven volunteers completed the online survey. One hundred and ten participants were students from a local university community in UK (86 females, 21 males, 3 non-binary; mean age = 21.58, SD = 4.18, range = 18–39), and 97 participants were from a cloud-sourcing platform (Prolific.co) that included volunteers from the US and UK (48 females, 49 males, 0 non-binary; mean age = 30.43, SD = 5.49, range = 19–40). The former group received research credits toward their psychology modules, and the latter group received £2.25, for participation. The criteria of recruitment included (1) being fluent in English, (2) being at the age between 18 and 40 years old, (3) having normal or corrected-to-normal visual acuity, and (4) having no color blindness. The age limit was justified by Tu et al.'s (2020) finding that older adults did not exhibit an uncanny valley effect. For the cloud-sourcing volunteers, they were also required to have an approval rate of higher than 95 % of at least 20 previous submissions, to ensure

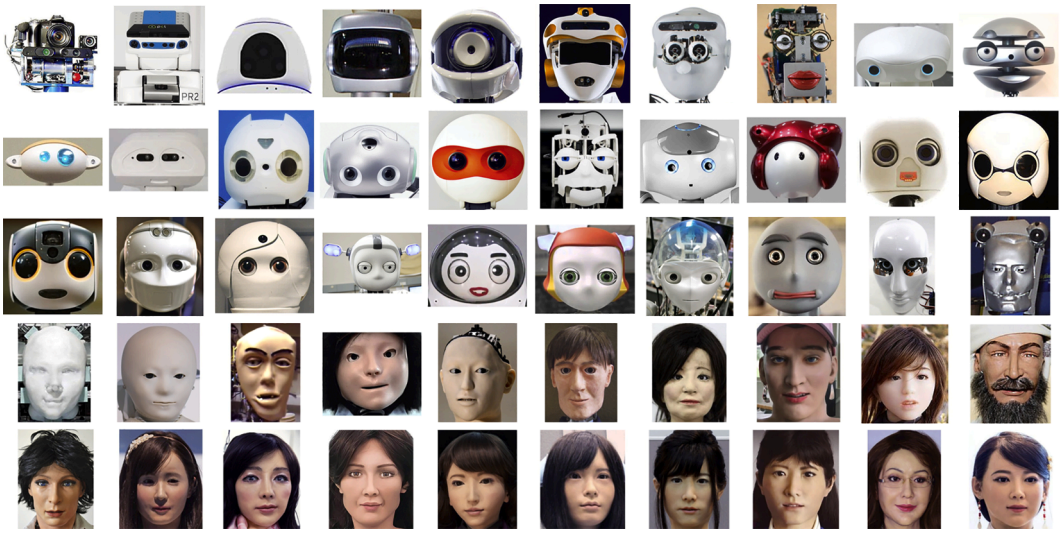

Fig. 2. Fifty robot faces selected from Experiment 1 and used in Experiments 2 and 3: from the top left to the bottom right, the robots are placed in the order of their MK scores (from low to high).

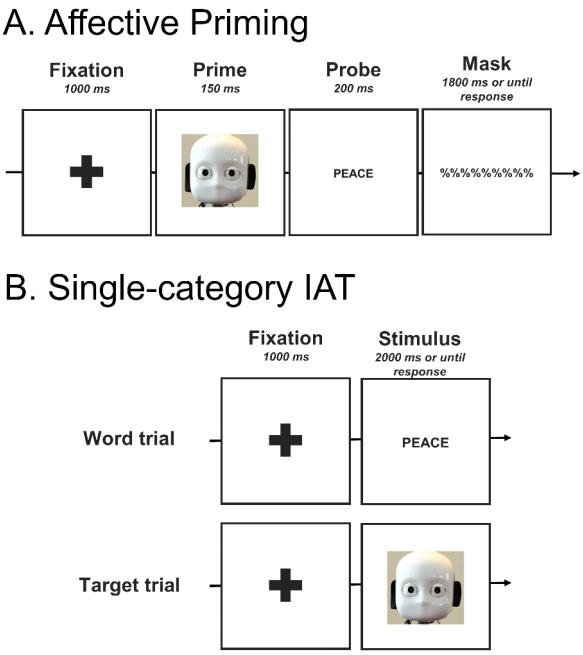

Fig. 3. Illustrations of the event sequence within a trial for the affective priming task in Experiment 2 (A) and the single-category IAT in Experiment 3 (B). The figures are for an illustrative purpose and are not scaled to the actual display size.

the quality of online data. The sample size was determined based on three previous studies (Mathur and Reichling, 2016; Mathur et al. 2020; Tu et al., 2020), which included sample sizes that varied between 66 and 358. The observed effect size was reported only in Tu et al.'s study. They reported a large effect size ($R^2 = 0.324$) for 77 young adults, which was the same age range as that of the present experiment. Hence, I first recruited as many participants from the local university community as I could in one academic term (Autumn 2022) and recruited more participants from the online platform in February 2023 to reach at least 200 participants in total. All participants provided informed consent; for those who did not consent, the survey was terminated automatically. The study was reviewed and approved by the Research Ethics Committee at the local university.

2.1.2. Materials

The online survey was created by using an online survey platform (Qualtrics), and all data were received and stored initially in their online server. The digital images of robot faces were those used by Mathur et al. (2020) and were retrieved from their OSF project page. Only the images of robots were included, excluding those of human faces. Robots' images were further removed if they were either duplicated or were similar to the other images in the stimulus set, which left 99 robot faces for the present study. The actual images used in the study can be found on the OSF project page for the present study (https://osf.io/3ae2k/).

2.1.3. Procedure

At the beginning of the online survey, participants were first presented with an array of all robot faces on a single page with the instructions that they were about to rate each of the 99 images of robot faces in two scales, (1) "how machine- or human-like the robot looks" (MH scale; −100 for "Extremely mechanical" and +100 for "Extremely human-like") and (2) "how pleasant or creepy it would be if you were to interact with them in an everyday situation such as asking questions in a museum information booth" (likability; −100 for "Less pleasant or creepy" and +100 for "More pleasant or enjoyable"). The phrases of these questions were based on those used by Mathur and Reichling's (2016) survey. Participants were then presented with one image at a time in a random order, which was determined for each participant by the survey tool (Qualtrics), and adjusted a slider for each question to give a score. The slider was initially positioned at 0 (the midpoint of the slider), and participants were not allowed to proceed to the next image until they moved the slider from the initial position. After they rated all images, they were given a brief explanation of the uncanny valley phenomenon and were asked whether they had experienced it in the past or during the survey. If they had, they were asked to describe why they thought some of the robots were repulsive or eerie in their own words.

2.1.4. Data analysis

The analysis script including all analyses carried out in the present study can be found on the OSF project page (https://osf.io/3ae2k/). The analyses were conducted and visualized in R Studio (R Core Team, 2021) with the following packages: lme4 (Bates, Maechler, Bolker, and Walker, 2015), tidyverse (Wickham et al., 2019), and ggpubr (Kassambara, 2022).

Following Mathur et al. (2020) and others (Mathur & Reichler, 2016; Tu et al., 2020), the data were analyzed with means by robot face as the

unit of analysis. Means and variances of MH and likability scores were calculated across participants for each face. Polynomial regression models were fit by regressing likability score on polynomial terms for MH score (e.g., MH, MH^2, MH^3, …). In the analysis, I fit second- to fifth-degree polynomial regression models with the data points weighted by the inverse of variance in the likability score at each corresponding MH score. To find the most parsimonious model, the fits of two models that differed in one degree were submitted to an F-test. That is, there were comparisons between the second- and third-degree models, between the third- and fourth-degree models, and the fourth- and fifth-degree models. The most parsimonious model was defined as the one that fit the data better than a lower degree model while a higher degree model did not improve the fit anymore (Mathur & Reichler, 2016). Mathur et al. used polynomial regression to mean-centered MH score, but the models were fit to raw scores and report the results below (see Fig. 4A); I also carried out the same analysis on standardized scores, which showed the same results (see Appendix).

The results of the present study were also compared to those of Mathur et al.'s by correlating standardized scores from the two studies. Because Murther et al. included images of human faces as well as those of robots, participants used higher MH ratings (> 0) predominantly for human faces and lower MH ratings (< 0) predominantly for robots, whereas participants in the present study were presented only with robot faces and used the entire scale to rate the same robots. Consequently, the same robot faces were rated less human-like in Mathur et al.'s study than in the present study. To minimize this context effect, the data from robot faces in Mathur et al.'s scores were standardized after removing those for human faces. The data from the present study was also standardized in the same manner. Note that the analysis was concerned with the correlational structure of the two sets of data, not the absolute differences, so that the standardization only minimized the context effect when the results were presented visually (as in Fig. 3) but should not alter the outcomes of the correlation analysis.

Table 1
Summary of model comparisons. The most parsimonious model was indicated in bold. All measures were regressed onto the polynomial terms of MH score. P(AP) is the positivity score from the affective priming procedure, and P(SC-IAT) is the positivity score from the single-category IAT. In Experiment 3, no model up to the fifth-degree model fit the RT data.

Study	Measure	Model Comparison	*F*	*p*	η_p^2
1	Likability	Second vs. third	F(1, 95) = 77.06	< 0.001	.394
		Third vs. fourth	**F(1, 94) = 25.23**	**< 0.001**	**.213**
		Fourth vs. fifth	F(1, 93) = 0.39	.535	.004
2	P(AP) for RT	**Second vs. third**	**F(1,46) = 5.83**	**.020**	**.111**
		Third vs. fourth	F(1,45) = 0.33	.569	.007
		Fourth vs. fifth	F(1,44) = 2.30	.137	.050
	P(AP) for PE	**Second vs. third**	**F(1, 46) = 6.22**	**.017**	**.118**
		Third vs. fourth	F(1,45) = 0.03	.873	.001
		Fourth vs. fifth	F(1,44) = 2.51	.120	.054
3	P(SC-IAT) for RT	Second vs. third	F(1, 46) = 0.63	.432	.013
		Third vs. fourth	F(1, 45) = 0.67	.417	.014
		Fourth vs. fifth	F(1, 44) = 2.72	.107	.058
	P(SC-IAT) for PE	Second vs. third	F(1, 46) = 9.54	.003	.159
		Third vs. fourth	**F(1, 45) = 6.39**	**.015**	**.127**
		Fourth vs. fifth	F(1, 44) = 0.04	.084	.001

2.2. Results

The results are summarized in Table 1. The fourth-degree polynomial satisfied the criteria for the most parsimonious model. As shown in Fig. 4A, the likability score initially increased as MH increased from –100, but it peaked at an MH score of around –60 and began to decrease thereafter. As MH score increased further, the likability score started to rise again at around an MH score of 10, which reached the bottom of the uncanny valley. To compare the results of the present study to Mathur et al.'s (2020) original study, the data from both studies are plotted in Fig. 3 (the same robot faces were connected by grey lines). It should be noted that ratings of eight images used in the present study were not included in the spreadsheet shared by Mathur et al.'s OSF page, and they were dropped from the correlational analysis. There were significantly positive correlations between the two studies for both MH score, Pearson's $r(N = 89) = 0.984$, $p < .001$, and likability score, Pearson's $r(N = 89) = 0.881$, $p < .001$. These results show that the uncanny valley effect was apparent in the present study and that the ratings for these robot faces were highly consistent with those of Mathur et al.'s.

2.3. Discussion

Experiment 1 showed the uncanny valley effect that was similar to

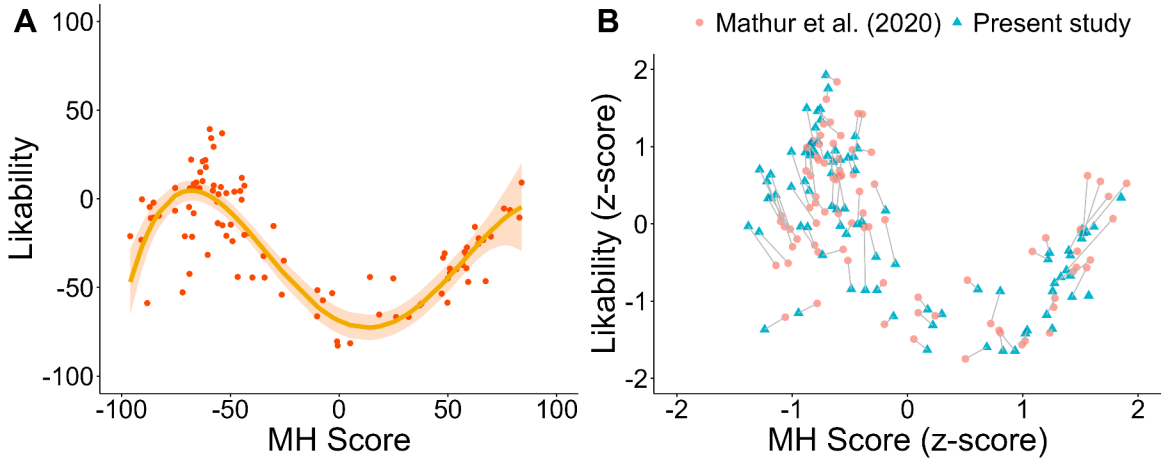

Fig. 4. Scatterplots of observed likability score on machine-humanness (MH) score in Experiment 1. (A) The solid line is the predictions of the fourth-degree polynomial regression models. The shaded band is the point-wise 95 % confidence interval. (B) Comparisons of the likability and MH scores for the respective robot faces between Mathur et al.'s (2020) study and the present study (the grey lines connect the scores from the same robot faces from the two studies).

the effect observed in Marthur et al.'s (2020) original study. It should be reminded that the comparison between the results of the present study to those of Marthur et al.'s was made based on scores that were standardized within the respective studies because Marthur et al.'s study included human faces and higher MH scores were used predominantly for human faces, whereas the present study only included robot faces and the entire MH scores were used to rate robots. Although the standardization of the rating scores reduced the context effect, it might not have eliminated it entirely. Experiment 1 obtained a correlation coefficient of 0.984 for MH scores and 0.881 for likability scores between the two studies, which means that 96.8 % of the variance in the MH score and 77.6 % of the variance in the likability score in one study were accounted for by the other. A portion of the remaining 3.2 % variance in the MH score and 22.4 % variance in the likability score may be attributed to such a residual context effect. Even so, however, the similarity between the two studies is rather impressive.

It should also be acknowledged that the present experiment was motivated by the wide variety of exemplars that were available in the stimulus set developed by Marthur et al. A drawback of this approach is that it lacks control over perceptual factors and image qualities across the range of stimuli (see Fig. 2). Therefore, although the uncanny valley effect was obtained along with the subjective rating of human-likeness (MK score), these stimuli do not permit us to infer the reason why the likability rating varies for these images or why the uncanny valley emerges for this particular stimulus set. Hence, more controlled stimuli are required to investigate perceptual and cognitive factors that elicit uncanny feelings in these stimuli.

Nonetheless, the successful replication of the uncanny valley effect in subjective rating provided a strong empirical basis to evaluate the two behavioral paradigms to measure affective reactions to the robot faces. The same robot faces as in Experiment 1 were used in these studies. Experiment 2 used the affective priming procedure (Fazio et al., 1986), and Experiment 3 used the single-category IAT (Karpinski and Steinman, 2006). To my knowledge, neither of these affective measures have been used to evaluate implicit responses to individual items. If the uncanny valley is being observed with these implicit affective measures, it would achieve two ends: First, it demonstrates that the uncanny valley does occur with implicit affective measures; this would indicate that the phenomenon could have impact on the way people behave toward these robots, not only on the feeling they might have toward the robots. Second, it demonstrates for the first time that these behavioural affective measures can be used to evaluate affective responses to individual items rather than categories of items. This could widen the use of these paradigms to address more practical questions beyond laboratory research.

3. Experiment 2

3.1. Method

3.1.1. Participants

There was no previous study that used an implicit measure to estimate the uncanny valley or that used an implicit measure at individual item levels, based on which sample size could be determined. Considering typical studies using cognitive experiments, I chose to recruit at least 50 participants in each of the two studies, which was a higher end of sample size typically used in such experiments with a fully within-subject design. Although the effect sizes for item-level implicit measures were still unknown, this sample size would be large enough to achieve a sufficient statistical power, given the large effect size (η_p^2 = 0.213 and 0.213) obtained for the subjective rating in Experiment 1; for example, to obtain a power of 0.9 or above, I would have needed 44 participants to obtain the effect size equal to $\eta_p^2 = 0.2$ in a repeated-measures F-test (Faul et al., 2009). At the end, there were 63 participants in Experiment 2. They were paid £10 for participation. All reported having normal or corrected-to-normal visual acuity, normal color vision, and being fluent in English. Participants were further screened for their overall accuracy in the affective priming task. Two participants marked an overall accuracy below 90 % and were excluded. The analyses were carried out with the remaining 61 participants (27 females, 34 males; mean age = 25.67, SD = 4.96, range = 18–42). The data including the excluded participants can be found on the OSF project page (https://osf.io/3ae2k/).

3.1.2. Apparatus and stimuli

Stimuli were selected from the robot faces validated in Experiment 1. The images were first sorted in an ascending order of their mean MH score obtained in Experiment 1, and 50 images at odd positions in this order were used for test trials. The remaining 49 images at even positions were used for practice trials. I decided to use only 50 images for test trials to keep a feasible duration of the session (< 30 min) while maintaining as many trials per image as possible. There were four positive words (*Happy, Cheer, Peace*, and *Love*) and four negative words (*Evil, Murder, Abuse*, and *Agony*). These words were used in previous IAT studies and had been shown to induce evaluative reactions (Greenwald et al., 1998; Yamaguchi and Beattie, 2020). The experiments were created with and was controlled by Inquisit 6 (Millisecond Software, LLC.), and the apparatus consisted of a personal computer and 24-in flat screen monitor.

3.1.3. Procedure

The experiment was conducted for a group of participants in a group testing room, in which five computers were arranged in each of the four rows. The number of participants in a session varied from 7 to 12. Each computer was partitioned from neighboring computers with screens, so that participants were not able to see or interact with others in the room once they were seated in front of their computers. Each session was attended by three or four experimenters who monitored possible disruptions during the session. Once seated, participants read on-screen instructions on the task, which emphasized the speed of responding while participants were also asked to keep their response accuracy above 90 %. A session took <45 min, including the time to seat and instruct participants as well as the time to complete the experiment.

The task was based on the design used by Eder, Leuthold, Rothermund, and Schweinberger (2012; see Fig. 3A). A trial started with a black fixation cross at the screen center on a white background for 1000 ms, followed by a prime that was one of the robot faces sampled randomly without replacement. The prime stayed on the screen for 150 ms and was replaced by a probe, which was one of the eight words. The probe was visible only for 200 ms and was masked with an array of nine "%" symbols. Participants had 2000 ms after the probe onset to make a response, which required them to press the "S" or "L" key on a QWERTY keyboard according to the valence of the word. The mapping between valence and key was randomly determined for each participant, with an equal probability of assigning the two keys to positive and negative meanings. Response time (RT) was the interval between an onset of the probe and a keypress. When a key was pressed, the trial ended immediately. If participants pressed an incorrect key or failed to make a response within 2000 ms, a red X was presented for 2000 ms, followed by a 500-ms blank display. If they pressed the correct key, they were presented only with the 500-ms blank display. The next trial started with the fixation cross.

Each participant first performed a block of 10 practice trials on categorizing words into positive and negative meanings. In this block, there was no mask after the probe, and the probe was visible until the trial ended. In the second block, also containing 10 practice trials, participants performed the same affective priming task, but the probe was visible only for 200 ms and was masked with an array of nine "%" symbols. The following four blocks consisted of 100 test trials each, and participants performed the same task as the second practice block. In each of these blocks, each of the 50 robot faces appeared twice, once with a positive word (positive trials) and once with a negative word (negative trials). In total, there were four positive trials and four

negative trials for each robot face for each participant.

After performing the affective priming task, participants rated each of the 50 robot faces on the MH scale as in Experiment 1. Each robot face appeared one at a time in a random order. Participants used a slider and adjusted the scale ranging from −100 to +100. The initial position was at a score of 0, and they were not permitted to proceed if the slider was not moved from the initial position. They were also asked to avoid only using the extreme values (−100 or +100) for all faces but to use the entire scale to rate robot faces. There was no timeout to make a response.

3.1.4. Data analysis

Mean RT for correct responses and percentages of error trials (PE) were computed for each participant, separately for positive trials and negative trials. Trials were discarded if RT was <200 ms (as immature responses) or if no response was made within the response deadline (2000 ms). Based on the logic behind affective priming, the more positively participants perceived the robot face, the faster (and the more accurate) the responses should be on positive trials but the slower (and the less accurate) they should be on negative trials. Therefore, for each robot face, RT and PE for positive trials was subtracted from RT and PE for negative trials to derive the *index of positivity* of participants' perception of the robot. Hence, the more positively the robot face was perceived, the higher the positivity score should be for the robot face. The resulting measure is denoted as P(AP), where P stands for positivity score.

The data were analyzed in the same manner as in Experiment 1, where the positivity score replaced the self-rating likability score. The positivity score was averaged across participants for each robot face and was polynomially regressed onto the MH score for the robot faces up to the fifth-degree model or until the most parsimonious model was found. As in Experiment 1, the most parsimonious model was that which fit the data better than a lower degree model while a higher degree model did not improve the fit significantly. This procedure was carried out with raw positivity and MH scores as well as the standardized scores, and the results of the standardized scores are reported in Appendix. The standardized scores were tested because the positivity and MH scores were in arbitrary scales, but the main results were largely consistent.

Although the original plan was to use the MH scores obtained after the affective priming task, these scores were much noisier than those obtained in Experiment 1. This was partly because several participants only used extreme values (−100 or +100) for almost all faces, despite the instructions not to do so. Also, all participants were already exposed to these images for a half hour and might not have reflected their first impressions. As Mathur et al. (2020) noted, these individual scores were a noisy version of the normative scores derived from the large sample in Experiment 1. Therefore, I decided to use the MH score from Experiment 1 for the present analyses so as to minimize the contamination from extraneous factors. Subjective ratings from Experiment 2 (and Experiment 3) are also available on the OSF project page for the purpose of reanalysis.

3.2. Results

Mean RT for positive trials was 609 ms (SE = 11.44), and that for

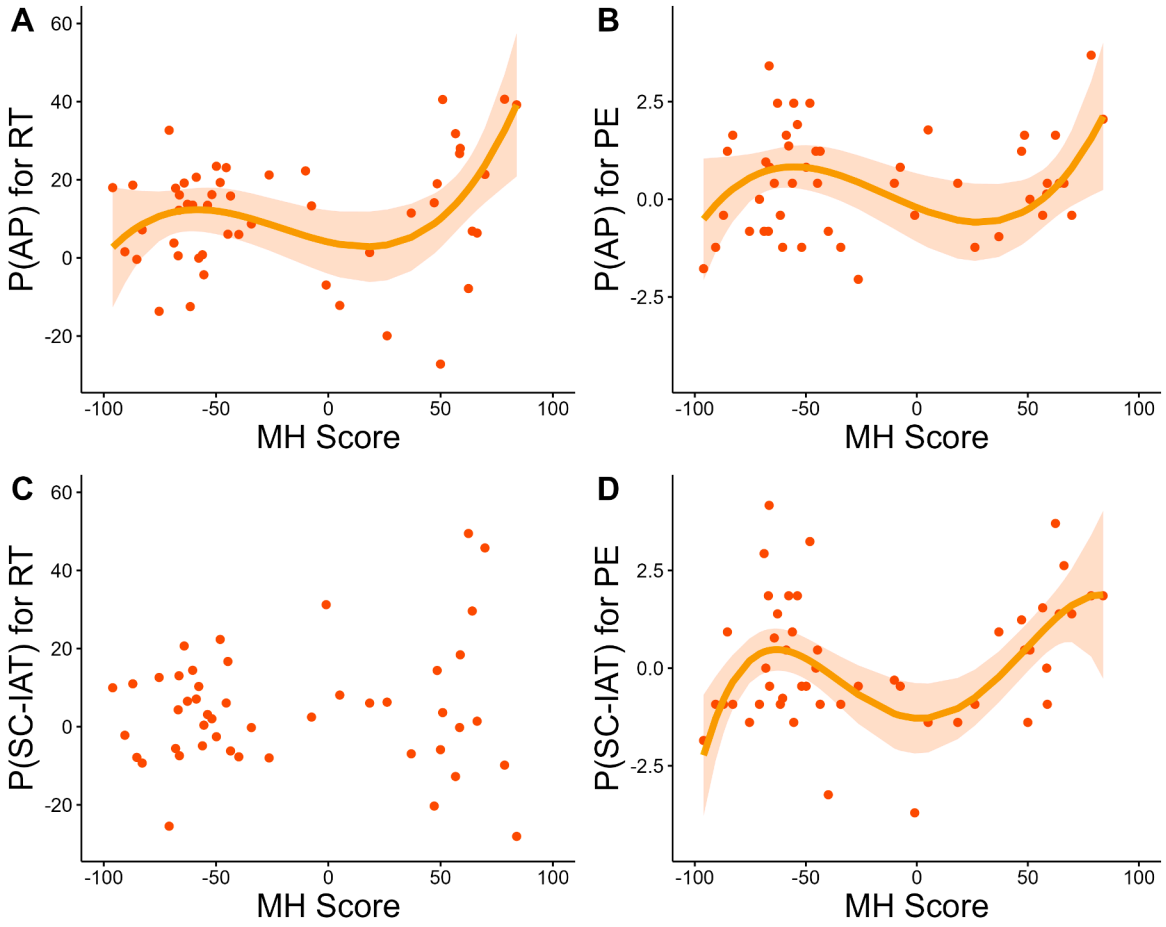

Fig. 5. The positivity scores for RT and PE in Experiment 2 and Experiment 3 regressed on polynomial terms of the machine-humanness (MH) score from Experiment 1. The solid lines are the predictions of the most parsimonious models for the respective measures. The shaded bands are the confidence intervals. The dots are mean scores for individual robot faces. P(AP) is the positivity score from the affective priming procedure in Experiment 2, and P(SC-IAT) is the positivity score from the single-category IAT in Experiment 3. No predictions are shown for P(SC-IAT) for RT because no model up to the fifth-degree polynomial fit the data.

negative trials was 620 ms (SE = 10.81). The former was significantly shorter than the latter, $t(60) = 3.08$, $p = .003$, Cohen's $d = 0.394$. PE for positive trials was 3.02 % (SE = 0.24), and that for negative trials was 3.39 % (SE = 0.33). There was no overall affective priming effect, $t(60) = 1.45$, $p = .152$, Cohen's $d = 0.186$. Thus, there was an overall positive affective priming effect in RT, but not in PE. Note that the lack of the affective priming effect in PE does not mean that there was no uncanny valley effect in the PE data, as shown below.

Next, the results of polynomial model comparisons are summarized in Table 1. For P(AP) for RT, the most parsimonious model was the third-degree polynomial, which is depicted in Fig. 5A. This model was significantly better than the second-degree model, whereas it was not significantly different from the fourth-degree model. For P(AP) for PE, the most parsimonious model was also the third-degree polynomial (see Fig. 5B). This model was significantly better than the second-degree model. Therefore, P(AP) for both RT and PE showed an increasing trend as the MH score increased, but it peaked at around –60 and began to decrease, consistent with the subjective likability rating in Experiment 1. The positivity measure then reached the bottom of a valley at an MH score between 20 and 40 and rose once again.

3.3. Discussion

Experiment 2 successfully reproduced the uncanny valley effect with the affective priming task, indicating that implicit affective reactions do follow the same pattern as the subjective rating of the likability of the robot faces. It is noteworthy that this result is obtained with a relatively small number of trials per item (4 positive and 4 negative; 8 trials in total), which is surprising because cognitive tasks typically involve many more trials per condition. The successful demonstration of the uncanny valley effect in the present experiment is likely because the data were analyzed based on the unit of robot faces, averaged across participants, rather than in terms of participants. In essence, data were averaged across all trials for 61 participants for each robot face, reducing the noise in the data analysis. Hence, the present analysis approach appears to be useful to evaluate item-level affective responses. The outcome indicates that the procedure is effective to evaluate affective reaction to individual items within a category at a group level analysis (i.e., when the scores were averaged across participants for each robot). Experiment 3 examined whether similar results could be obtained with the single-category IAT.

4. Experiment 3

4.1. Method

4.1.1. Participants

Fifty nine participants were recruited from the same subject pool as in Experiment 2. Data collection for Experiment 3 was carried out after completing that for Experiment 2, and participants were not permitted to participate in both studies. Five participants were excluded for low overall response accuracy, and the analysis was carried out for the remaining 54 participants (24 females, 28 males, 2 non-binary; mean age = 27.26, SD = 5.06, range = 18–37).

4.1.2. Apparatus, stimuli, and procedure

The stimuli were the same digital images of robots used in Experiment 2. Positive words were *Happy, Cheer, Peace, Love*, and *Pleasure*; negative words were *Evil, Murder, Abuse, Agony*, and *Filth*.[c] The same apparatus and test room as in Experiment 2 were used.

The task was the single-category IAT based on the design originally used by Karpinski and Steinman (2006; see Fig. 3B). A trial started with a black fixation cross, against a white background, for 1000 ms, followed by an imperative stimulus (either a word or an image of a robot face). The imperative stimulus stayed on the screen until a response was made or for 2000 ms if no response was made. If an incorrect response was made or no response was made, a red "X" appeared on the display for 2000 ms, followed by a 500-ms blank display. If the correct response was made, participants were presented only with the blank display. The next trial started with the fixation cross. RT was the interval between an onset of the imperative stimulus and a keypress. Each participant started with a block of 10 practice trials for which they were presented only with a word on each trial and categorized it as a positive or negative word by pressing the "S" or "L" key. As in Experiment 2, the mapping of positive and negative words to the two keys was randomly determined for each participant. In the following two blocks of 10 practice trials each, this word categorization task was mixed with the target task for which participants were presented with a robot face and pressed one of the two keys. In one of the blocks, they pressed the key assigned to positive words (positive block), and in the other block, they pressed the key assigned to negative words (negative block). The word trial appeared with the probability of 0.29 (10 out of 35) and the target trial appeared with the probability of 0.71 (25 out of 35). The order of these blocks was randomly determined for each participant and kept fixed for the remaining blocks for the participant. After the practice blocks, there were 16 blocks of 35 test trials each, which consisted of 10 trials for the word task and 25 trials for the target trials. Each of the 10 words occurred once per block. For a set of four consecutive blocks, each face occurred on 2 trials, once in the positive block and once in the negative block. In total, each face occurred on four positive trials and four negative trials for each participant. Therefore, there were 400 face trials (50 faces x 8 trials per face) and 160 word trials (10 words x 16 blocks). After performing the single-category IAT, participants also rated each of the 50 robot faces on the MH scale as in Experiment 2.

4.1.3. Data analysis

Mean RT for correct responses and PE were computed for each participant. Trials were discarded if RT was <200 ms (as immature responses) or if no response was made within the response deadline (2000 ms). As the IAT effect was evaluated for each robot face, only data from the target task was used. As in Experiment 2, for each robot face, RT and PE for positive trials was subtracted from RT and PE for negative trials to derive the index of positivity of participants' perception for the robot, now labeled as P(SC-IAT). The positivity score was polynomially regressed onto the MH score from Experiment 1.

4.2. Results

Mean RT was 521 ms (SE = 9.40) for positive trials and 523 ms (SE = 8.60) for negative trials, and there was no significant difference between these types of trial, $t(53) = 0.39$, $p = .701$, Cohen's $d = 0.053$. PE was 1.98 % (SE = 0.33) for positive trials and 2.19 % (SE = 0.32) for negative trials, which were again not significantly different, $t(53) = 0.63$, $p = .533$, Cohen's $d = 0.085$. Hence, there was no overall single-category

[c] Experiment 3 used 10 words (5 positive and 5 negative), instead of 8 words as in Experiment 2, because our previous study on the IAT (Yamaguchi & Beattie, 2019) used the same word set. Note that the length of a single trial was slightly longer in Experiment 2 than in Experiment 3 because there were two stimuli per trial (prime and probe) in Experiment 2, instead of a single stimulus (picture or word) in Experiment 3. Hence, to shorten the length of a session in Experiment 2, we used 8 words instead of 10 words.

IAT effect in RT or PE. Note again that the lack of the single-category IAT effect in the overall descriptive statistics does not necessarily indicate a lack of the uncanny valley effect in these data, which was examined in the polynomial regression as reported below.

Unlike the results of the affective priming task in Experiment 2, there was no parsimonious model up to the six degree polynomial for P(SC-IAT) for RT (see Fig. 5C). However, P(SC-IAT) for PE still produced the uncanny valley effect (see Fig. 5D). The most parsimonious model was the fourth-degree polynomial, which was significantly better than the third-degree model but was not significantly different from the fifth-degree model for both raw and standardized scores. The positivity score increased as HM score increased from –100 and peaked at around –70. It then dropped to the bottom of the valley at a HM score of zero and then increased toward more positive values for higher HM scores. Thus, although P(SC-IAT) for RT did not show the uncanny valley effect, P(SC-IAT) for PE still showed the expected pattern that was consistent with the results of Experiment 1.

4.3. Discussion

The present experiment used the single-category IAT, which again demonstrated the uncanny valley effect in PE but not in RT. The discrepancy between RT and PE is sometimes obtained in the IAT, especially when participants did not have sufficient time to make speeded responses (as confirmed in an unpublished study manipulating speed-accuracy tradeoffs in another experiment from our lab). The response deadline of 2000 ms might have been too short for many participants in the single-category IAT as it required switching between the two tasks in a random order. A recent study also suggested that RT in the IAT only accounted for a small portion of variance in predicting explicit attitude measures (Schimmack, 2021), which may imply that the RT measure in the IAT might not be sensitive enough to capture the variability of affective reactions across robot faces. Nevertheless, the uncanny valley effect obtained in PE indicates that the single-category IAT is still useful to evaluate affective reactions to individual items.

5. General discussion

Fifty years after Mori (1970) proposed the original hypothesis, the relevance of the uncanny valley has become more evident in recent years, owing to the current expansion of robotics and AI technologies as well as the popularity of computer graphics animations in films and digital media. Understanding the emotional impact of these technologies and media is still an underrepresented area of research, and the present study filled a caveat in the study of the uncanny valley effect by demonstrating the effectiveness of behavioral paradigms to measure affective reactions to individual items. Despite the wide acknowledgement, the uncanny valley remained to be a controversial hypothesis because of the mixed support it has received from empirical studies. For example, Bartneck and Kanda (2009) compared people's likability ratings for a human and for an android of the same human and found no significant difference, based on which they argued that the uncanny valley should no longer be used to hold back the development of highly realistic androids. Burleigh et al. (2013) also argued against the uncanny valley effect, suggesting that the phenomenon depended on multiple different factors that gave rise to the human-likeness of robots or avatars. In contrast, several recent studies observed the uncanny valley effect successfully by using a large sample of existing robot faces (e.g., Mathur et al. 2020; Tu et al., 2020), based on which the present study was designed. Possible reasons for the mixed support stem from the ambiguities in the original formulation of the uncanny valley. One of the ambiguities comes from the original Japanese term "*Shinwakan*" used to describe the affinity for robots. This is a composite of two terms that cannot be translated readily into other languages. In fact, different researchers have adopted different translations of the term, and the presence of the uncanny valley effect apparently depended on which translation was used, especially when self-rating is used to measure this attribute (e.g., Ho and MacDorman, 2016; MacDorman and Chattopadhyay, 2016). The use of a behavioral measure in the present study is, thus, a step forward to overcome this ambiguity in the uncanny valley effect.

The present behavioral measures are also free from the assumption about the underlying causes of the phenomenon. Several theories have been proposed to explain the uncanny valley effect; some researchers attribute the phenomenon to perceptual properties such as the difficulty of discrimination (human and robot exemplars may be difficult to discriminate when they appear similar to each other; Cheetham et al., 2014) or category ambiguity (exemplars are categorically ambiguous when they lay at the boundary between two categories; Yamada et al., 201), whereas others attribute it to cognitive processes such as cognitive conflict (Ferrey et al., 2015; Weis and Wiese, 2017) and mind attribution (Appel et al., 2020; Gray & Wegner, 2012). Depending on their theoretical positions, different measures of affinity have been used to test the uncanny valley effect. The present behavioral measures are relatively free from these assumptions because they are designed to measure affective responses directly and, arguably, implicitly. The present findings of the uncanny valley in the two behavioral measures, thus, assure a degree of confidence that the effect has an emotional basis. In the meantime, the interpretations of the outcomes of the two behavioral measures are still subject to debate, which is considered in more detail later.

In addition to the importance of the present results for the study of the uncanny valley, the present findings are also important as they demonstrated that the behavioral paradigms provided item-level measures of affective reactions. The positivity scores from the affective priming procedure and the single-category IAT showed a non-linear relationship with human-likeness of robot faces similar to the uncanny valley effect observed in the self-rating likability scale replicated in Experiment 1. This result demonstrated for the first time that these behavioral measures can be used to evaluate affective responses to individual items rather than to groups of items. The sensitivity to affective values of individual items is an encouraging finding that suggests a wider applicability of the paradigms in many other domains outside laboratory research, such as marketing and economic research or product development where implicit evaluations of individuals products within a category (e.g., different models of a car brand) are more meaningful than that of categories of products (cars vs. motor cycles). Because this is the first demonstration of item-level affective measures, further scrutiny is required to evaluate the robustness of the present findings and the generalizability of the approach to other stimuli.

While emphasizing the usefulness of the behavioral measures, some pitfall should also be acknowledged. Although the two behavioral paradigms used in the present study were the arguably most popular implicit measures of affective reactions (e.g., Bar-Anan and Nosek, 2014), their results were not in perfect agreement. The affective priming procedure produced the uncanny valley effect similarly in the RT and PE measures, whereas the single-category IAT only produced the effect in PE. The lack of the effect in RT could be due to a potential speed-accuracy trade-off; in a study from our lab that is yet to be published, the IAT effect emerged only in PE when response speed was emphasized, whereas it emerged only in RT when response accuracy was emphasized. However, it could also be due to the unreliability of the IAT measure that has been raised by several researchers, especially for its standard format. There has been an extensive debate as to the validity of the IAT as a pure measure of implicit attitude (e.g., Blanton et al., 2009; Fiedler, Messner, and Bluemke, 2006; Karpinski and Hilton, 2001; Ziegert et al., 2009; Yamaguchi & Beattie, 2019), and the debate is still ongoing (Cvencek et al., 2020; Schimmack, 2021). Considering the volume and intensity of the debate, it is no longer possible to make a decisive claim based on a single study. Yet, the fact that the two behavioral measures yielded similar uncanny valley effects at the item-level analyses in the present study indicates that both procedures

do share a certain commonality of the underlying affective reactions to these robot faces.

At the same time, however, the IAT's failure to produce the uncanny valley effect in RT also underscores the difference between them. Researchers of social evaluation have highlighted two loci of implicit biases; one that arises from affective reactions (e.g., fear of robots) and the other that arises from stereotypes (e.g., robots are unintelligent). As much as affective reactions do, stereotypes can also yield implicit biases in human's behavior and judgment, but they are semantic instead of emotional. While different variations of the IAT have been developed to tease apart the two sources of implicit biases and have been shown to account for different types of human behavior (Amodio and Devine, 2006), it is still possible that the IAT and the affective priming used in the present study rely on these loci of implicit biases to different degrees, such that affective priming reflected affective reactions more than the IAT, or vice versa. Although it is difficult to examine how much of the two behavioral measures reflects affective reactions and how much reflects stereotypes, the present results seem to imply that affective priming might be a better option in terms of the robustness (because the uncanny valley effect was observed both in RT and in PE) and to avoid the ambiguity in the interpretations inherent in the IAT. To corroborate this, it is worth emphasizing that the IAT effect is usually measured in RT rather than in PE, so the lack of the uncanny valley effect in RT raises a concern.

5.1. Does it scale up to real-time interactions with robots?

Most studies of the uncanny valley used static images to evaluate human reactions to robots and avatars, but humanoid robots are designed to interact with human users in real time and in a dynamic fashion. A salient feature of real-time interaction is the movement of a robot. Mori (1970) originally proposed that the motion would amplify the uncanny valley, which should result in stronger negative reactions to robots. In line with this proposal, several studies have demonstrated the importance of robots' movement in human perception of the robots. Moriguchi et al. (2010) asked participants to rate the animacy of a human, a robot, and an android in video clips in which they performed a card-sorting task. The movements of the robot and android were more mechanical than that of the human actor, and they found significantly lower animacy scores for these artificial agents than that for the human actor. It was also found that human observers had a tendency to imitate the movement of a humanoid robot (Oztop, Chaminade, and Franklin, 2005) but not that of an industrial robot (Milner et al., 2003). More subtle bodily motions are also found to alter the human perception of robots. Noma et al. (2006) tested whether human observers identified an android as a human when it was static (sitting on a chair) or when it was making natural movements of the eyes and chest. They found that only 20 % of the participants identified it as a human in the static condition, but 76.9 % identified it as a human when the android robot exhibited the natural movements of the eyes and chest. Similarly, several studies found that human actors are sensitive to the gaze (e.g., Belkaid et al., 2021; Boucher et al., 2012) and head movement (Sidner et al., 2005) of a robot partner in a cooperative or conversational task. These subtle non-verbal behavioral cues contribute to human users' likability of the robots (Zinina, Zaidelman, Arinkin, & Kotov, 2020). Nevertheless, empirical evidence supporting Mori's suggestion that motion amplifies the uncanny valley has not been found (Piwek, McKay, & Pollick, 2014).

The human-likeness of a robot is unlikely to rely on a single factor; instead, it is a multidimensional construct that involves interplay between different factors (Burleigh et al., 2013; MacDorman, 2006). Hence, a systematic manipulation of human-likeness is extremely difficult, and the results have been variable across different studies that used different exemplars and manipulations. Indeed, Hanson (2006) argued that the uncanniness of androids is nothing to do with human-likeness, but it depends on other aesthetic elements of the face. One way to overcome this multidimensional nature of human-likeness is to use a wide range of exemplars, as was done in the present study, rather than distorting or morphing a small sample of images. Although the repeated replications of the uncanny valley effect with this approach make it safe to conclude that the uncanny valley is robust and generalizable across different participant populations, they do not rule out all possible confounding factors that underlie the human-likeness of robots, nor do they guarantee that the same result is obtained with a different set of exemplars. There is always a possibility that the sample is biased in one way or the other (e.g., there could be a bias in the type of robot engineers produce). A shortcoming of the use of existing robots as exemplars is that it does not allow us to understand what is responsible for the observed uncanny valley effect. Even if the effect is robust, we are left with uncertainties as to what is causing the affective reactions observed in the present study. Another shortcoming of this approach is that it relies on self-rating of human-likeness by observers, which could be noisy and subject to different interpretations. Hence, a more objective measure of the human-likeness scale needs to be explored in future studies as well. In fact, to my knowledge, no study of the uncanny valley effect has posed the question of what underlies the perception of human-likeness of non-human characters (but see Diel & Lewis, 2024). This is yet another unexplored realm of the field that requires future investigations.

Robot technologies have been advancing quickly in the last decade, and it is not rare to see movies of working humanoid robots on social media or video sharing platforms. Applications of behavioral measures in a dynamic interactive situation involving robots can be challenging, but it is still possible. For example, the aforementioned studies by Boucher et al. (2012) and by Belkaid et al. (2021) demonstrated innovative approaches to examine the influence of a robot's gaze on human cognitive performance. In these studies, human participants sat face-to-face with a robot partner and performed a cooperative task for which the robot's gaze behavior was systematically manipulated. The results of these studies clearly showed that the influences of the robot's gaze behavior can be observed in human reaction time. The approaches taken by these researchers also point to a possible extension of the present behavioral approach to examine human affective reactions to different robot designs in a collaborative task setting . Yet, the use of real robots necessarily limits the researcher's ability to compare a large number of exemplars that would be necessary to observe the uncanny valley. Instead, HCI researchers might wish to start their investigation with static images, as in the present study, or with short video clips to present multimodal information of robots (e.g., Grundke, Stein, and Appel, 2023; MacDorman, 2006; Mitchell et al., 2011). They can then narrow down to a smaller number of exemplar robots to test affective reactions in a real-time interaction. Whether findings in the uncanny valley research scale up to a real-time interaction with a robot is yet another exciting and important topic of future research.

Finally, the use of behavioral methods like the ones in the present study may seem to demand more unique skills, more time, and more training, to administer them properly, as compared to online questionnaires, only because they are less familiar to many researchers outside the field. However, the procedure could be standardized, and data processing can be pipelined, as the benefit of the approach is established firmly in the future investigations, which should ease the use of the behavioral methods for wider applications outside laboratories. It is noteworthy that the IAT has been used as an online tool for many years (e.g., Project Implicit; https://www.projectimplicit.net/). Therefore, there are fewer barriers than one would imagine in making use of behavioral methods, and it is indeed an interesting avenue for future investigations as to whether more simplified procedures than the ones used in the present study could provide reliable affective measures outside laboratories.

5.2. Concluding remarks

The ever expanding markets of smart AI technologies raise the importance of considering their emotional impact on human users,

which would play a key role in people's acceptance of the technologies in the next few years. The robustness of the uncanny valley effect found in the present study suggests the possibility that the phenomenon originates from intrinsic human reactions to human-like figures that still lack the perfection of a real human appearance. Provided the complex array of possible factors that come into play in creating the eventual feeling toward these agents, further investigations are needed to understand the underlying mechanism of the uncanny valley. An interesting approach is to develop a computational model to simulate the uncanny valley (Moore, 2012), which has been used to simulate hypothetical differences in the perception of artificial agents between the general population and individuals with an autism spectrum disorder (Ueyama, 2015). Although interesting an approach to evaluate individual differences of the uncanny valley, the empirical validity of the computational model is yet to be established.

It is also acknowledged that the polynomial model used in the present study is useful to describe the relationship between the human-likeness (MK) score and the observed subjective or behavioral measures of likability, but it does not explain the underlying psychological processes. Diel and Lewis (2024) published a new study challenging the validity of a polynomial model as an 'explanation' of the uncanny valley and proposed an alternative approach, which consists of a combination of multiple linear models that fit to different categories of faces separately. They suggested that "the uncanny valley can be thought of as a moderated linear function" (p. 7) in which sensitivity to distortion is high at higher human-likeness levels, increasing relative uncanniness for human-like faces. That is, they suggested that the uncanny valley can be explained by two factors; facial categories that vary in realism and configural distortions of faces within each category. While this suggestion is insightful, an application of their framework to an arbitrary set of facial exemplars (such as those used in the present study) is difficult, because there is no easy way to classify exemplars into pre-defined categories with a measurable distortion level. Nevertheless, the authors introduced a novel approach to interpret the uncanny valley based on a psychologically interpretable model. More psychological models of the uncanny valley effect are needed to advance our understanding of this phenomenon.

At the present, there are more unknowns in our understanding of the underlying mechanism of the uncanny valley. Given the fast rate of development in robotics and AI technologies with a human level performance, it may only be a matter of time when humans spend more time interacting with artificial agents than with real humans (e.g., imagine how much time students might be using ChatGPT nowadays, as opposed to vising their teacher's office hours, to prepare for an incoming exam). Understanding the psychological impact of human-robot interaction can become a key factor to enhance well-being of the humankind in the near future.

CRediT authorship contribution statement

Motonori Yamaguchi: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author declares the following financial interests/personal relationships which may be considered as potential competing interests:

Motonori Yamaguchi reports financial support was provided by ESSEXlab (University of Essex).

Supplementary materials

Appendix can be found, in the online version, at doi:10.1016/j.ijhcs.2024.103443

Data availability

The data, analysis scripts, and experimental programs are available at the Open Science Framework project page (https://osf.io/3ae2k/).

References

Amodio, D.M., Devine, P.G., 2006. Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. J. Pers. Soc. Psychol. 91, 652–661.

Appel, M., Izydorczyk, D., Weber, S., Mara, M., Lischetzke, T., 2020. The uncanny of mind in a machine: humanoid robots as tools, agents, and experiencers. Comput. Human Behav. 102, 274–286.

Bar-Anan, Y., Nosek, B.A., 2014. A comparative investigation of seven indirect attitude measures. Behav. Res. Methods 46, 668–688.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bartneck, C., Kanda, T., 2009. My robotic doppelgänger—A critical look at the uncanny valley. In: Proceedings of the 18thIEEE International Symposium on Robot and Human Interactive Communication, pp. 269–276.

Belkaid, M., Kompatsiari, K., De Tommaso, D., Zblith, I., Wykowska, A., 2021. Mutual gaze with a robot affects human neural activity and delay decision-making processes. Sci. Robot. 6, eabc5044.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., Tetlock, P.E., 2009. Strong claims and weak evidence: reassessing the predictive validity of the IAT. J. Appl. Psychol. 94, 567–582.

Bluemke, M., Friese, M., 2008. Reliability and validity of the single-target IAT (ST-IAT): assessing automatic affect towards multiple attitude objects. Eur. J. Soc. Psychol. 38, 977–997.

Boucher, J.-D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., Dominey, P.F., Ventre-Dominey, J., 2012. I reach faster when I see you look: gaze effects in human-human and human-robot face-to-face cooperation. Front. Neurorobt. 6, 3.

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. Comput. Human Behav. 29, 759–771.

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., Jancke, L., 2013. Category processing and the human likeness dimension of the uncanny valley hypothesis: eye-tracking data. Front. Psychol. 4, 108. Article.

Cheetham, M., Suter, P., Jäncke, L., 2011. The human likeness dimension of the "uncanny valley hypothesis": behavioral and functional MRI findings. Front. Hum Neurosci. 5, 126. Article.

Cheetham, M., Suter, P., Jancke, L., 2014. Perceptual discrimination difficulty and familiarity in the uncanny valley: more like a "happy valley. Front. Psychol. 5, 1219. Article.

Cvencek, D., Meltzoff, A.N., Maddox, C.D., Nosek, B.A., Rudman, L.A., Devos, T., Dunham, Y., Baron, A.S., Steffens, M.C., Lane, K., Horcajo, J., Ashburn-Nardo, L., Quinby, A., Srivastava, S.B., Schmidt, K., Aidman, E., Tang, E., Farnham, S., Mellott, D.S., Banaji, M.R., Greenwald, A.G, 2020. Meta-analytic use of balanced identity theory to validate the implicit association test. Personal. Soc. Psychol. Bull. 47, 185–200.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., Moors, A., 2009. Implicit measures: a normative analysis and review. Psychol. Bull. 135, 347–368.

Diel, A., Lewis, M., 2024. Rethinking the uncanny valley as a moderated linear function: Perceptual specialization increases the uncanniness of facial distortions. Computers in Human Behavior 157, 108254. https://doi.org/10.1016/j.chb.2024.108254.

Diel, A., MacDorman, K.F., 2021. Creepy cats and strange high houses: support for configural processing in testing predictions of nine uncanny valley theories. J. Vis. 21, 1–20.

Diel, A., Weigelt, S., MacDorman, K.F., 2021. A meta-analysis of the uncanny valley's independent and dependent variables. ACM Trans. Human-Robot Int. 11, 1. Article.

Eder, A.B., Leuthold, H., Rothermund, K., Schweinberger, S.R., 2012. Automatic response activation in sequential affective priming: an ERP study. Soc. Cognitive Affect. Neurosci. 7, 436–445.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G., 2009. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav. Res. Methods 41, 1149–1160.

Fazio, R.H., Sanbonmatsu, D.M., Powell, M.C., Kardes, F.R., 1986. On the automatic activation of attitudes. J. Pers. Soc. Psychol. 50, 229–238.

Ferrey, A.E., Burleigh, T.J., Fenske, M.J., 2015. Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. Front. Psychol. 6, 249. Article.

Fiedler, K., Messner, C., Bluemke, M., 2006. Unresolved problems with the "I", the "A", and the "T": a logical and psychometric critique of the implicit association test (IAT). Eur. Rev. Soc. Psychol. 17, 74–147.

Gray, K., Wegner, D.M., 2012. Feeling robots and human zombies: mind perception and the uncanny valley. Cognition 125 (1), 125–130.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: the implicit association test. J. Pers. Soc. Psychol. 74, 1464–1480.

Grundke, A., Stein, J.-P., Appel, M., 2023. Improving evaluations of advanced robots by depicting them in harmful situations. Comput. Human Behav. 140, 107565.

Hanson, D., 2006. Exploring the aesthetic range from humanoid robots. In: Proceedings of the ICCS/CogSci-2006 Long Symposium: Toward social mechanisms of android science. Vancouver, Canada, pp. 16–20.

Ho, C.-C., MacDorman, K.F., 2016. Measuring the uncanny valley effect: refinements to indices for perceived humanness, attractiveness, and eeriness. Int. J. Soc. Robot. 9, 129–139.
Karpinski, A., Hilton, J.L., 2001. Attitudes and the Implicit Association Test. J. Pers. Soc. Psychol. 81, 774–788.
Karpinski, A., Steinman, R.B., 2006. The single category implicit association test as a measure of implicit social cognition. J. Pers. Soc. Psychol. 91, 16–32.
Kassambara, A. (2022). ggpubr: 'ggplot2' based publication ready plots. R package version 0.5.0. https://CRAN.R-project.org/package=ggpubr.
Kätsyri, J., Frörger, 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. Article.
Kilner, J.M., Paulignan, Y., Blakemore, S.J., 2003. An interference effect of observed biological movement on action. Curr. Biol. 13, 522–525.
Lischetzke, T., Izydorczyk, D., Hüller, C., Appel, M., 2017. The topography of the uncanny valley and individuals' need for structure: a nonlinear mixed effect analysis. J. Res. Pers. 68, 96–113.
MacDorman, K.F., 2006. Subjective rating of robot video clips for human likeness, familiarity, and eeriness: an exploration of the uncanny valley. In: Proceedings of ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science. Vancouver, Canada, pp. 26–29.
MacDorman, K.F., Chattopadhyay, D., 2016. Reducing consistency in human realism increases the uncanny valley effect: increasing category uncertainty does not. Cognition 146, 190–205.
MacDorman, K.F., Entezari, S.O., 2015. Individual differences predict sensitivity to the uncanny valley. Interac. Stud 16, 141–173.
MacDorman, K.F., Vasudevan, S.K., Ho, C.-C., 2009a. Does Japan really have robot mania? COMPARING attitudes by implicit and explicit measures. AI Soc. 23, 485–510.
MacDorman, K.F., Ishiguro, H., 2006. The uncanny advantage of using androids in cognitive and social science research. Interact. Stud. 7, 297–337.
MacDorman, K.F., Green, R.D., Ho, C.-C., Koch, C.T., 2009b. Too real for comfort? Uncanny responses to computer generated faces. Comput. Human Behav. 25, 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. M., Levitan, C.A., Nave, G., Manfredi, D., Bessette-Symons, B., Szuts, A., Balazs, A., 2020. Uncanny but not confusing: multisitestudy of perceptual category confusion in the uncanny valley. Comput. Human Behav. 103, 21–30.
Matsuda, Y., Okamoto, Y., Ida, M., Okanoya, K., Myowa-Yamakoshi, M., 2012. Infants prefer the faces of strangers or mothers to morphed faces: an uncanny valley between social novelty and familiarity. Biol. Lett. 8, 725–728.
Minato, T., Shimada, M., Itakura, S., Lee, K., Ishiguro, H., 2006. Evaluating the human likeness of an android by comparing gaze behaviors elicited by the android and a person. Adv. Robot. 20, 1147–1163.
Mitchell, W.J., Szerszen Sr, K.A., Lu, A.S., Schermerhorn, P.W., Scheutz, M., MacDorman, K.F, 2011. A mismatch in the human realism of face and voice produces and uncanny valley. Iperception 2, 10–12.
Moore, R.K., 2012. A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. Sci. Rep. 2, 864.
Mori, M., 1970. Uncanny Valley. Energy 33–35.
Mori, M., 2012. The uncanny valley (K. F., MacDorman, & N. Kageki, Trans.). IEEE Robot. Automat. Magazine 98–100. Original work published 1970.
Moriguchi, Y., Minato, T., Ishiguro, H., Shinohara, I., Itakura, S., 2010. Cues that trigger social transmission of disinhibition in young children. J. Exp. Child Psychol. 107, 181–187.
Noma, M., Saiwaki, N., Itakura, S., Ishiguro, H., 2006. Composition and evaluation of the humanlike motions of an android. In: Proceedings of the 6thIEEE-RAS International Conference on Humanoid Robots, pp. 163–168.
O'Shea, B.A., Wiers, R.W., 2020. Moving beyond the relative assessment of implicit biases: navigating the complexities of absolute measurement. Soc. Cogn. 38, S187–S207.
Oztop, E., Chaminade, T., Franklin, D.W., 2005. Human-humanoid interaction: is a humanoid robot perceived as a human? Int. J. Humanoid Rob. 2, 537–559.
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, 271–277.
R Core Team, 2021. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Schimmack, U., 2021. The implicit association test: a method in search of a construct. Perspect. Psychol. Sci. 16, 396–414.
Seyama, J., Nagayama, R., 2007. The uncanny valley: effect of realism on the impression of artificial human faces. Presence 16, 337–351.
Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C., 2005. Explorations in engagement for humans and robots. Artif. Intell. 166, 140–164.
Strait, M.K., Floerke, V.A., Ju, W., Maddox, K., Remedios, J.D., Jung, M.F., Urry, H.L., 2017. Understanding the uncanny: both atypical features and category ambiguity provoke aversion toward humanlike robots. Front Psychol. 8, 1366. Article.
Tinwell, A., Abdel Nabi, D., Charlton, J.P., 2013. Perception of psychopathy and the uncanny valley in virtual characters. Comput. Human Behav. 29, 1617–1625.
Tu, Y.-C., Chien, S.-E., Yeh, S.-L., 2020. Age-related differences in the uncanny valley effect. Gerontology 66, 382–392.
Ueyama, Y., 2015. A Bayesian model of the uncanny valley effect for explaining the effects of therapeutic robots in autism spectrum disorder. PLoS ONE 10, e0138642.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.
Wang, S., Lilienfeld, S.O., Rochat, P., 2015. The uncanny valley: existence and explanations. Rev. Gen. Psychol. 19, 393–407.
Weis, P.P., Wiese, E., 2017. Cognitive conflict as possible origin of the uncanny valley. In: Proceddings of the Human Factors and Ergonomics Society 2017 Annual Meeting, pp. 1599–1603.
Yamada, Y., Kawabe, T., Ihaya, K., 2013. Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. Japan. Psychol. Res. 55, 20–32.
Yamaguchi, M., Beattie, G., 2020. The role of explicit categorization in the implicit association test. J. Experiment. Psychol. General 149, 809–827.
Ziegert, J.C., Hanges, P., 2009. Strong rebuttal for weak criticisms: reply to Blanton et al. (2009). J. Appl. Psychol. 94, 590–597.
Zinina, A., Zaidelman, L., Arinkin, N., Kotov, A., 2020. Non-verbal behavior of the robot companion: a contribution to the likeability. Procedia Comput. Sci. 169, 800–806.