

Virtual Faces Evoke Only a Weak Uncanny Valley Effect: An Empirical Investigation With Controlled Virtual Face Images

Perception

0(0) 1–24

© The Author(s) 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0301006619869134

journals.sagepub.com/home/pec**Jari Kätsyri** 

Department of Cognitive Neuroscience, Maastricht University, the Netherlands; Department of Computer Science, Aalto University, Finland

Beatrice de Gelder

Department of Cognitive Neuroscience, Maastricht University, the Netherlands

Tapio Takala 

Department of Computer Science, Aalto University, Finland

Abstract

The uncanny valley (UV) hypothesis suggests that increasingly human-like robots or virtual characters elicit more familiarity in their observers (positive affinity) with the exception of near-human characters that elicit strong feelings of eeriness (negative affinity). We studied this hypothesis in three experiments with carefully matched images of virtual faces varying from artificial to realistic. We investigated both painted and computer-generated (CG) faces to tap a broad range of human-likeness and to test whether CG faces would be particularly sensitive to the UV effect. Overall, we observed a linear relationship with a slight upward curvature between human-likeness and affinity. In other words, less realistic faces triggered greater eeriness in an accelerating manner. We also observed a weak UV effect for CG faces; however, least human-like faces elicited much more negative affinity in comparison. We conclude that although CG faces elicit a weak UV effect, this effect is not fully analogous to the original UV hypothesis. Instead, the subjective evaluation curve for face images resembles an *uncanny slope* more than a UV. Based on our results, we also argue that subjective affinity should be contrasted against subjective rather than objective measures of human-likeness when testing UV.

Corresponding author:

Jari Kätsyri, Department of Computer Science, Aalto University, P.O. Box 15400, FI-00076 Aalto, Finland.

Email: jari.katsyri@aalto.fi

Keywords

face perception, animacy perception, uncanny valley hypothesis, social cognition

Date Received: 26 April 2019; accepted: 20 July 2019

Introduction

Imagine you are watching a film depicting a boy trapped on a raft in the middle of the ocean with only a Bengal tiger as company. As you watch the story unfold, you at times get an inexplicable sense of unease when observing the tiger's appearance and movements. Only later, you learn that the tiger was computer-animated rather than real, which may have contributed to your feelings of unease. This example of a possible reaction from the film *Life of Pi* (Lee, 2012) illustrates a hypothesis called the uncanny valley (UV; Mori, 1970/2012). This hypothesis, as originally suggested for humanoid robots, suggests that entities that appear near-human can elicit negative subjective reactions in their observers. Following the most recent translation of Mori's (1970/2012) original article written in Japanese, we use here the term *affinity* to refer to the range of subjective experiences associated with the UV, ranging from eeriness (Jpn. *bukimi*) to familiarity (*shin-wakan*). Eeriness in particular has been identified as a hallmark of subjective uncanniness in recent conceptual and empirical work (e.g., Chattopadhyay & MacDorman, 2016; Ho & MacDorman, 2017; Mangan, 2015). Mori explicitly recommended that robot designers should aim for only modestly human-like appearing robots in order to avoid the UV (see Kageki, 2012), and similar recommendations have been given more recently for creating animated film characters (e.g., Butler & Joschko, 2009). This design strategy is not always possible, however. In particular, highly realistic computer-generated (CG) faces are now often used to study emotions and social cognition in humans (for some examples, see Balas & Pacella, 2015; Krumhuber, Tamarit, Roesch, & Scherer, 2012; Marschner, Pannasch, Schulz, & Graupner, 2015; Schilbach et al., 2006). Here, we aim to investigate whether and to what extent the UV hypothesis poses an obstacle to exploiting near-human-like CG faces in this kind of research.

In Figure 1, we propose three different shapes for the relationship between human-likeness and affinity. The UV hypothesis predicts that when affinity is plotted against human-likeness, an initial positive peak occurs at intermediate human-likeness levels, a negative peak occurs at high human-likeness levels, and the curve again reaches positive affinity at its terminus at complete human-likeness (as in Figure 1(b) and (c)). Empirical evidence for this kind of evaluative curve has remained surprisingly elusive, however (for reviews, see Kätsyri, Förger, Mäkärräinen, & Takala, 2015; Wang, Lilienfeld, & Rochat, 2015). On the contrary, the bulk of empirical evidence seems to favor a linear relationship or, in the present terminology, an *uncanny slope* between affinity and human-likeness (Figure 1(a)), in which decreasing levels of human-likeness are associated with increasing levels of eeriness (Kätsyri et al., 2015).

We suggest a further distinction between weak and strong variants of the UV in Figure 1 (b) and (c). Both of these variants predict that the UV effect occurs at high human-likeness levels, but their predictions differ with respect to whether other levels are allowed to evoke even greater eeriness in comparison. Unlike the strong variant, the hypothesis of a weak UV allows that some lower human-likeness levels may evoke greater negative affinity than the UV (Figure 1(b)). This kind of result pattern has been reported previously. For example,

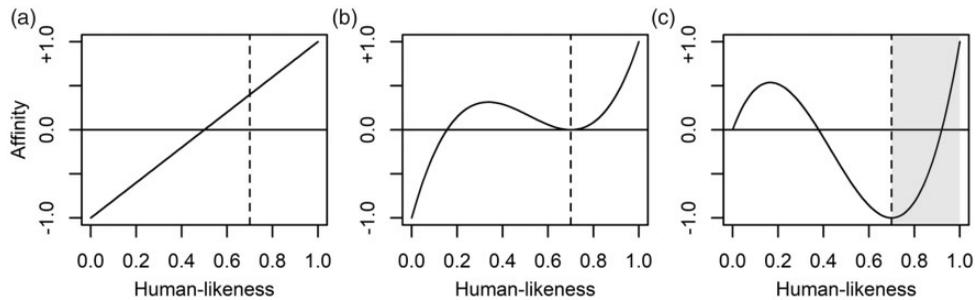


Figure 1. Hypothetical uncanny curves illustrated with polynomial functions. (a) A positive linear relationship or, in our terminology, an *uncanny slope* between human-likeness and affinity. (b) A *weak uncanny valley* with a negative affinity peak that does not constitute the lowest affinity overall. (c) A *strong uncanny valley* with a negative affinity peak that clearly elicits the lowest affinity. Dashed line illustrates the theoretically predicted location of the uncanny valley (70% human, cf. Weis & Wiese, 2017). The shaded region in Panel (c) illustrates the uncanny curve falling on the right side of the uncanny valley (e.g., for a continuum ranging from CG to human faces).

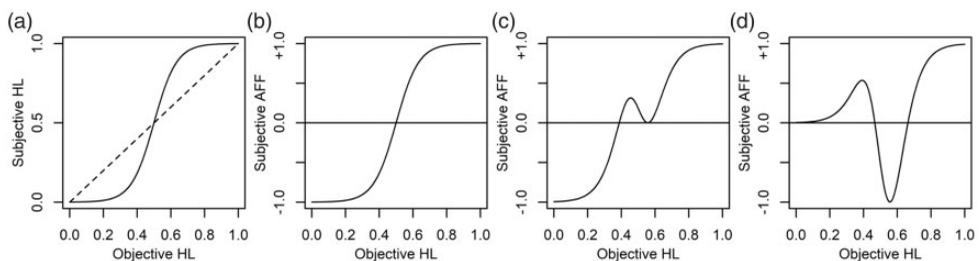


Figure 2. Hypothetical categorical perception effects on the uncanny valley. (a) A logistic curve (with $x_0 = 0.5$ and $k = 50$) for subjective versus objective HL. Dashed line illustrates a linear relationship for comparison. The same logistic curve for subjective AFF versus objective HL is shown in Panels (b) to (d) assuming that the relationship between subjective HL and AFF followed (b) an uncanny slope, (c) a weak uncanny valley, or (d) a strong uncanny valley effect (Figure 1(a) to (c)). HL = human-likeness; AFF = affinity.

MacDorman and Chattopadhyay (2016) showed that while CG faces with inconsistent realism levels elicited a valley-like effect for evaluated affinity, the least realistic CG faces still elicited more negative affinity in comparison. Similarly, Carr, Hofree, Sheldon, Saygin, and Winkielman (2017) reported that even though an android robot elicited more negative evaluations than a real human in terms of approachability, likability, and weirdness, its mechanical variant always elicited more negative or at best equally high evaluations (see results for absolute ratings in Carr et al., 2017). In strong UV, the lowest affinity levels always occur at the negative affinity peak (Figure 1(c)). Trivially, strong UV has more severe practical consequences than the weak UV. Previous theoretical postulations also seem to support the notion of a strong UV. First, this kind of curve is consistent with the original UV hypothesis (see Figures 1 and 2 in Mori, 1970/2012). Second, as noted by Moore (2012), a full explanation of the UV needs to explain why it evokes negative affinity and not just a lack of familiarity (see Figures 2 and 3 in Moore, 2012). Third, many of the concepts traditionally associated with the UV—such as eeriness, corpses, and zombies (Mori, 1970/2012)—imply the presence of extreme negative emotions in the observer.

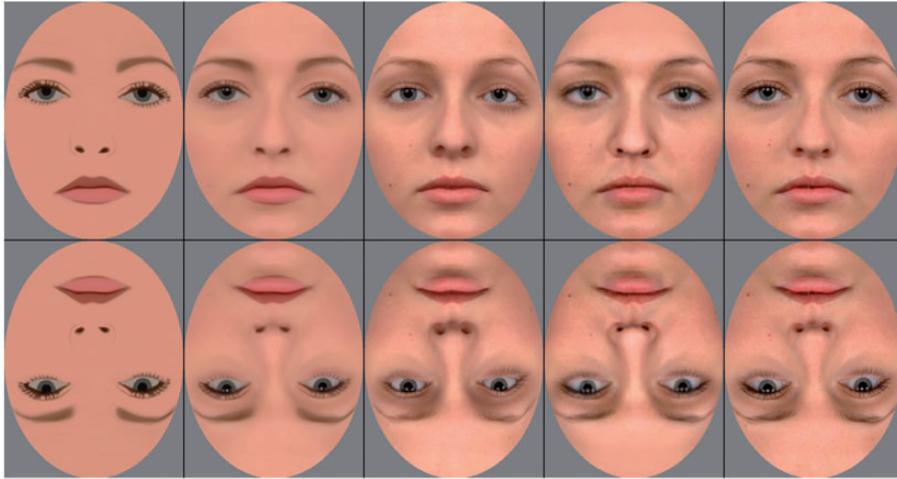


Figure 3. Sample upright (upper row) and inverted (bottom row) face images from Experiment 1. From left to right, the faces depict painted.simple, painted.shaded, CG.MakeHuman, CG.FaceGen, and human face types.

Results from several studies using naturalistic stimuli (uncontrolled stimuli adopted from real-life contexts) have provided potential support for the strong UV (e.g., Kätsyri, Mäkäräinen, & Takala, 2017; Lischetzke, Izydorczyk, Hüller, & Appel, 2017; MacDorman & Ishiguro, 2006; Mathur & Reichling, 2016; McDonnell, Breidt, & Bühlhoff, 2012; Piwek, McKay, & Pollick, 2014; Poliakoff, Beach, Best, Howard, & Gowen, 2013; Schindler, Zell, Botsch, & Kissler, 2017; Wang & Rochat, 2017; Yamada, Kawabe, & Ihaya, 2013). Examples of previously used naturalistic stimuli include images of robot faces (Mathur & Reichling, 2016), images of hand prostheses (Poliakoff et al., 2013), varying face images (Wang & Rochat, 2017), animation films (Kätsyri et al., 2017), and different CG rendering methods applied to faces (McDonnell et al., 2012; Schindler et al., 2017) and bodies (Carter, Mahler, & Hodgins, 2013; Piwek et al., 2014). Naturalistic and controlled stimuli possess contrasting advantages and disadvantages for studying the UV. An argument in favor of naturalistic stimuli is that because the causes of the UV are still not completely understood, naturalistic stimuli are more appropriate for tapping this real-life phenomenon than rigorously controlled experimental stimuli. An argument against naturalistic stimuli is that they may contain confound factors that cannot be fully controlled either by a priori stimulus selection or a posteriori statistical measures. As argued elsewhere (Kätsyri et al., 2015), including purposefully ill or morbid characters (e.g., zombie faces) or purposefully neonatal characters (e.g., toy-like robots or cartoon faces), both can confound the pure effects of human-likeness on affinity. There are many other potential confounds including but not limited to image quality (e.g., image compression artifacts, varying brightness and color conditions), social cues (emotional facial expressions, gaze and head directions, gender, attractiveness, and implied personality), design aesthetics, and observer-dependent factors (previous familiarity with the observed characters, expertise on the stimulus domain [e.g., animation films]).

Controlled stimulus generation methods offer much greater control over confound factors (but at the potential cost of decreased ecological validity). Image morphing is arguably one of the most commonly applied controlled stimulus generation methods for investigating the UV. This method has been used, for example, to create image continua from artificial to

human-like faces where the artificial images are CG faces (Cheetham, Suter, & Jäncke, 2011; MacDorman & Chattopadhyay, 2016), robot faces (Lischetzke et al., 2017; MacDorman & Ishiguro, 2006), cartoon faces (Sasaki, Ihaya, & Yamada, 2017; Yamada et al., 2013), or doll faces (Looser & Wheatley, 2010; Seyama & Nagayama, 2009). Image morphing is a particularly promising method for studying the UV because it allows generating well-controlled continua with several intermediate steps. Nevertheless, image morphing seems to suffer from two problems. On the one hand, when image morphing is carried out for dissimilar source and target images (e.g., for cartoon and human faces), ghosting artifacts are likely to occur where features present in only one of the images remain partly visible in the intermediate morphs. Such artifacts can appear eerie by themselves and confound any uncanny effects. On the other hand, using similar source images for morphing (e.g., CG and human faces) may severely restrict the generated range of human-likeness. As illustrated in the shaded region of Figure 1(c), if the left side of the morphed continuum fell into the UV, one might mistakenly conclude that the relationship between affinity and human-likeness resembles an uncanny slope rather than a UV. Using only CG faces as the starting point may be particularly problematic, given that CG faces tend to elicit negative affinity because their individual facial features represent inconsistent realism levels (MacDorman & Chattopadhyay, 2017).

Finally, we consider the issue of whether subjective affinity should be plotted against subjective or objective measures of human-likeness. With objective measures, we refer to quantitative human-likeness manipulation levels such as morph percentages between artificial and real faces. Several recent studies support the notion that artificial and human faces are perceived categorically (Cheetham et al., 2011; Looser & Wheatley, 2010; MacDorman & Chattopadhyay, 2016; Moore, 2012). As illustrated in Figure 2(a), this means that the relationship between subjective and objective human-likeness levels should follow an S-shaped logistic curve rather than a straight linear curve. Assuming that subjective evaluations would follow an uncanny slope (Figure 1(a)), subjective affinity would then also follow a similar S-shaped pattern when plotted against objective human-likeness, as illustrated in Figure 2(b). This likely explains why some previous studies (e.g., Thompson, Trafton, & McKnight, 2011) observed a poor fit when trying to fit polynomial functions to the relationship between subjective affinity and objective human-likeness. As shown in Figure 2(c), the logistic relationship between subjective and objective human-likeness would also distort the weak uncanny effect (Figure 1(b)) if subjective affinity was plotted against objective rather than subjective human-likeness. A further argument against objective human-likeness is that the same objective human-likeness scale cannot be used as a common metric for different human-likeness manipulations. For example, *percentage human-like* scale would not be meaningful when comparing morphed continua that began from different source images (e.g., painted and CG faces).

Although we acknowledge that naturalistic and controlled stimuli have distinct advantages and disadvantages for studying the UV, we have opted here for controlled stimuli as a more conservative approach. To the best of our knowledge, strong UV has not yet been demonstrated using rigorously controlled and unproblematic stimuli. We have identified several potential problems for typical controlled stimuli: morphing artifacts caused by too dissimilar source images, narrow human-likeness ranges caused by too similar source images, and the use of objective rather than subjective measures of human-likeness. In the current three studies, we aimed to avoid these potential problems by using image morphing with rigorously controlled source images. We used painted images of faces in addition to CG faces to tap a broader range of human-likeness than in previous studies. Importantly, we used painted, CG, and original variants of the same faces to minimize

differences between the source images. Finally, we compared continua beginning from CG and painted faces to test the prediction that CG faces would be more sensitive to the uncanny effect than other artificial faces.

Experiment 1: Painted, CG, and Human Faces

In our first study, we compared CG faces categorically against painted and human faces. Painted faces were intended to be minimally human-like stimuli whose features could still be matched rigorously with the original faces. Our main prediction was that CG faces would evoke more negative affinity than either painted or human faces (H1a), which would then provide support for the strong uncanny effect (Figure 1(c)) for CG faces. Based on uncanny slope and weak UV effects (Figure 1(a) and (b)), our alternative hypothesis was that CG faces would elicit more negative affinity than human faces but more positive affinity than painted faces (H1b).

As a secondary research question, we investigated the effects of inversion on the evaluation of painted, CG, and human faces. Inversion is considered a hallmark of perceptual expertise, and it typically manifests itself as the much slower and less accurate processing of inverted (rotated 180°) as compared with upright faces (Maurer, Le Grand, & Mondloch, 2002). Previous evidence suggests that inversion also affects human-likeness judgments such that it makes human faces more difficult to recognize as human, but it does not affect the recognition of artificial faces (Fan et al., 2014). More specifically, inversion elicits decreased human-likeness and increased eeriness (Almaraz, 2017) as well as less frequent attributions of mind (Deska, Almaraz, & Hugenberg, 2017) for faces residing on the right side of the category boundary that separates human from artificial faces. Given that face inversion has a greater impact on configural rather than featural processing (Maurer et al., 2002), these results have been interpreted to mean that the perception of humanness depends critically on the integration of facial features into a unified Gestalt (Hugenberg et al., 2016). Based on these findings, we predicted that inversion would elicit lower human-likeness and higher eeriness for human faces but not for painted or CG faces (H2).

Methods

Ethics. All studies (Experiments 1–3) were performed in accordance with the Declaration of Helsinki and approved by the Ethical Review Committee, Psychology and Neuroscience, Maastricht University (approval no. ERCPN-170_02_08_2016).

Participants. Participants were 36 (24 females) university students with a mean age of 19.9 years (standard deviation [*SD*] = 1.5 years). Two additional participants failed to pass diagnostic tests (see “Procedure” section in Experiment 1) and were excluded from the present data. Participants signed up to the study anonymously using the SONA system (<http://www.sona-systems.com>) of Maastricht University and received course credit in compensation for their participation. All participants provided informed consent prior to the beginning of the experiment.

Stimuli. Research stimuli are illustrated in Figure 3. Initial stimuli were frontal neutral-face images from 12 actors (6 females) in the Radboud (Langner et al., 2010) face image set (Identifiers 1, 2, 5, 8, 9, 24, 30, 32, 36, 37, 58, and 71). We generated two variants of both painted and CG faces to increase the generalizability of our results. Professional computer artist (R. B.) created the painted face variants using Adobe Photoshop software (Version

CS6). A simple painted face (painted.simple) shows a flat two-dimensional image with painted eye brow, eye, nose, and mouth regions aligned with the originals. A shaded painted face (painted.shaded) variant additionally contained shading cues, which were derived by applying the Photoshop oil paint filter to the original images. The computer artist also generated the first variant of CG faces (CG.MakeHuman) using MakeHuman (Version 1.1.1; <http://www.makehumancommunity.org/>) and Blender software (Version 2.7.9; <http://www.blender.org>). Specifically, MakeHuman software was used to derive base facial shape from the original image, which was then exported to Blender for asymmetry and other adjustments and the superimposition of facial texture. Second variants of CG faces (CG.FaceGen) were created using FaceGen Modeller (Version 3.13; Singular Inversions). Frontal and side images were imported into FaceGen, and initial alignment was provided manually using a number of feature points. Reconstructed faces were matched to originals with respect to scaling and head orientation. Most obvious artifacts in FaceGen images were corrected using Photoshop: Black line between the lips was removed, color errors in eyes and nostrils were corrected, and dark shadows next to the nose were removed by adjusting intensity histograms within the nose region (see Supplement).

All images were oval masked to conceal ears and hair. Oval regions in painted and CG faces were then matched to human faces in MATLAB (Version R2016a). Painted faces were matched with respect to mean pixel intensity values, and CG faces were matched with respect to both means and *SDs*. Additional adjustments to eye and mouth regions were done when necessary. Matching was carried out individually for each actor. Following previous conventions (e.g., Kobayashi, Otsuka, Kanazawa, Yamaguchi, & Kakigi, 2012; Railo, Karhu, Mast, Pesonen, & Koivisto, 2016), matching was carried out in RGB color space separately for each channel. Size for final images was 246×328 pixels.

Design. We used a 5 (face type: painted.simple, painted.shaded, CG.MakeHuman, CG.FaceGen, human) \times 2 (orientation: upright, inverted) within-subjects design. Orientation was nested within two counterbalanced stimulus sets such that participants in the same set always saw each specific actor in the same upright or inverted orientation. Participants were assigned randomly into sets.

Procedure. This study was carried out as an online evaluation, which was programmed and hosted in the Qualtrics platform (<http://www.qualtrics.com>). Only participants using a laptop or a desktop computer with a sufficiently large display (minimum 12") were accepted into the study. Sixty stimuli (5 face variants \times 2 orientations \times 6 actors per face set) were presented in a pseudorandomized order. Participants were asked to evaluate the human-likeness and eeriness of each face using three items. Human-likeness index included the items inanimate–living, artificial–realistic, and human-made–human-like (Cronbach's $\alpha = .95^1$). Eeriness index included the items typical–strange, familiar–eerie, and unusual–usual (reverse-coded; $\alpha = .87$). These items were adapted from Ho and MacDorman (2010, 2017); however, only three high-loading items were included from original indices, and the items were rephrased to fit the present context (cf. MacDorman & Chattopadhyay, 2016). Items were rated on a 7-step semantic differential scale ranging from -3 to 3 (e.g., $-3 = \textit{very artificial}$ to $3 = \textit{very realistic}$), and these ratings were averaged for scale summary. For consistency with the original UV hypothesis, eeriness scale was reversed to form a continuum from negative to positive affinity (from high to low eeriness). Participants were instructed to carry out the questionnaire at their own pace and in a single session without breaks. We used two diagnostic tests to identify careless responders: We presented one *Thatcherized* face (a face with mouth and eye regions inverted) among other stimuli,

and we asked participants to report whether they had answered all questions seriously. Participants who either failed to provide above zero eeriness ratings for the diagnostic face or who answered *no* to the latter question were excluded.

Results and Discussion

Figure 4 illustrates human-likeness and affinity evaluations for the different types of faces presented in upright and inverted orientations. Evaluation data were analyzed in R (Version 3.5.2; R Core Team, 2016) with packages *afex* and *emmeans* using a mixed-design analysis of variance with Greenhouse–Geisser correction for nonsphericity when appropriate. Face type and orientation were defined as within-subjects variables and stimulus set as a between-subjects variable of no interest. First, we note that face type had a significant effect on human-likeness, $F(2.28, 77.39) = 324.67$, $p < .001$, $\eta_G^2 = .83$, with significant pairwise differences between all face types (Bonferroni-corrected $p_{\text{Bonf.}} < .001$). This means that even the present highly realistic CG faces could be easily recognized as nonhuman. On the other hand, painted faces received much lower human-likeness ratings than CG faces (Figure 4(a)), which confirms the suggestion that relying only on CG faces would tap a narrow range of human-likeness. As can be seen in Figure 4(b), affinity results clearly favored hypothesis H1b over H1a. That is, the most negative affinity was elicited by painted and not by CG faces. Face type consistently exerted a statistically significant effect on eeriness ratings, $F(1.82, 62.02) = 66.07$, $p < .001$, $\eta_G^2 = .50$, and except for a nonsignificant difference between MakeHuman and FaceGen CG face variants ($p_{\text{Bonf.}} > .999$), all pairwise differences between face types were statistically significant ($p_{\text{Bonf.}} < .003$). These findings hence provided support against the strong UV for CG faces.

In our second hypothesis, we predicted that inversion would elicit lower human-likeness and more negative affinity but only for human faces. We observed a significant two-way interaction between face type and orientation for both human-likeness ratings, $F(3.07, 104.41) = 12.76$, $p < .001$, $\eta_G^2 = .04$, and eeriness ratings, $F(3.44, 117.12) = 3.67$, $p = .011$, $\eta_G^2 = .01$. Post hoc comparisons showed that, as expected based on previous studies (Almaraz, 2017; Deska et al., 2017), inverted human faces elicited lower human-likeness

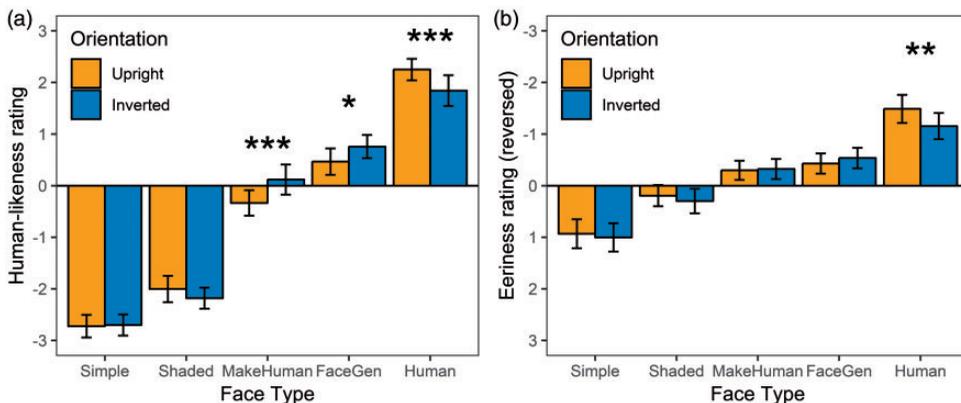


Figure 4. (a) Human-likeness and (b) affinity evaluations for painted (simple and shaded), computer-generated (MakeHuman and FaceGen), and human faces. Affinity evaluations refer to reverse-coded eeriness ratings (higher values denote lesser eeriness). Error bars denote within-subjects 95% confidence intervals (Morey, 2008), and asterisks denote statistically significant differences (* $p < .05$, ** $p < .01$, *** $p < .001$).

($p_{\text{Bonf.}} < .001$) and higher eeriness ratings ($p_{\text{Bonf.}} = .002$) than upright human faces. However, inversion also elicited significantly higher human-likeness ratings for both variants of the CG faces (MakeHuman: $p_{\text{Bonf.}} < .001$; FaceGen: $p_{\text{Bonf.}} = .010$). Inversion did not significantly affect human-likeness ratings for painted faces ($p_{\text{Bonf.}} > .117$) or eeriness ratings for either painted or CG faces ($p_{\text{Bonf.}} > .318$). Hence, these findings demonstrate that inversion impairs not only the recognition of human faces as human but also the recognition of CG faces as nonhuman, plausibly because inversion impairs the integration of facial features in both of them. This suggests that CG faces are special in the sense that they are processed in a unified rather than piecemeal fashion.

Experiment 2: Painted–Human and CG–Human Face Continua

Even though our first experiment provided support against the strong UV hypothesis, it is nevertheless still possible that a more fine-grained human-likeness continuum would reveal considerably different results. In particular, strong UV effect could occur at other human-likeness levels than that represented by CG faces. Furthermore, the first experiment was not designed to differentiate between uncanny slope and weak UV effects. In our second experiment, we therefore investigated two gradual continua: from painted to human faces (painted continuum) and from CG to human faces (CG continuum). We first considered the relationship between subjective and objective evaluations. In our first hypothesis (H1), we predicted that both subjective human-likeness and subjective affinity would follow a logistic (Figure 2(a) and (b)) rather than a cubic (third degree) polynomial curve (Figure 1(b) and (c)) when plotted against objective human-likeness. Cubic polynomials were used here because they were expected to capture most of the plausible UV shapes.

Second, we considered the overall relationship between subjective affinity and subjective human-likeness. Given that previous empirical evidence provides more support for the uncanny slope rather than the UV effect (Kätsyri et al., 2015), we predicted that affinity and human-likeness would show a positive linear relationship with each other (H2a). As an alternative hypothesis, we predicted that this relationship would resemble a weak UV (Figure 1(c)). Specifically, we predicted that the relationship between human-likeness and affinity would follow a cubic polynomial curve in which a local affinity minimum occurred at high levels of human-likeness (H2b).

Finally, we considered differences between CG and painted continua. Recent evidence suggests that the UV effect can be elicited by faces possessing realism-inconsistent features (MacDorman & Chattopadhyay, 2016; Seyama & Nagayama, 2009). It has also been suggested that realism inconsistency is characteristic of CG faces (MacDorman & Chattopadhyay, 2017). Assuming that CG faces would indeed appear more eerie than equally human-like faces on the painted continuum, the evaluation curve for CG faces should begin from a location that is below the evaluation curve for painted faces on the affinity axis. Consequently, CG curve should rise more steeply toward the human end point than the painted curve. Hence, we predicted that the evaluation curve for subjective affinity versus subjective human-likeness would be steeper for CG than for painted continuum (H3).

Methods

Participants. Participants were 44 (34 females; $M = 20.0$ years, $SD = 3.4$ years) university students. Two additional participants were excluded because of lack of variance in the affinity ratings and because of clearly deviant affinity ratings for painted faces (i.e., opposite to the

majority), respectively. All participants provided informed consent prior to the beginning of the experiment. Participants received course credit in compensation for their participation.

Stimuli. Initial stimuli were the same as in Experiment 1 except for three changes. First, we used only one variant of painted (simple painted) and CG (MakeHuman) faces. Second, we dropped one male and one female actor (ids. 8 and 24) to reduce fatigue, which left us with 10 actors. Third, corneal light reflections were adapted from original images and added to painted faces to avoid partly transparent corneal reflections in intermediate morphs. For the final stimuli, we created nine intermediate images (12.5% morph step) in painted and CG continua. For painted continuum, this was accomplished in Photoshop by placing a human image layer on top of a painted image layer and adjusting its opacity from 0% to 100%. For CG continuum, we used FantaMorph software (Version 5.4.7; Abrosoft) due to small differences between the MakeHuman-generated and original images. Final stimulus samples are illustrated in Figure 5.

Design. The study used a 2 continuum: painted, CG) \times 9 (human-likeness level) within-subjects design. Human end point, which was common to both continua, was presented only once.

Procedure. This experiment was carried out as a laboratory study. After arrival, participants received an introduction to the experiment and signed an informed consent form. Each participant evaluated the human-likeness and affinity of all stimuli in two separate blocks, with the block order counterbalanced across participants. To reduce fatigue, we used single items for human-likeness and eeriness, which were adapted from the study of Mathur and Reichling (2016). Specifically, participants were asked to rate human-likeness and affinity using Visual Analogue Scales whose values were coded from -100 to 100 . Human-likeness ranged from *extremely artificial* to *extremely human-like*, and affinity ranged from *extremely unpleasant and creepy* to *extremely pleasant and not at all creepy*. These terms were only explained before their corresponding evaluation blocks to reduce anticipatory effects. Blocks were separated by a 2-minute break. There were six practice stimuli that represented middle and end points on the painted and CG continua, and which came from two actors not included in the actual study. Each evaluated face remained on the screen until participant had given his or her response. Human end point, which was identical for painted and CG continua, was shown only once; hence, participants rated a total of 170 stimuli (9 painted and 8 CG levels*10 actors) in both blocks. Stimuli were presented in a

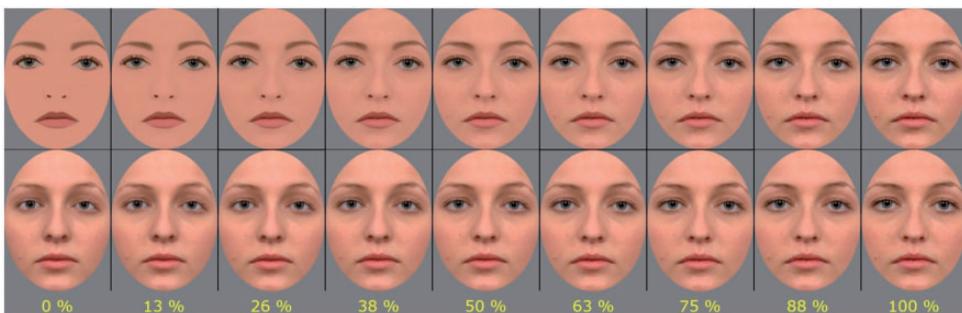


Figure 5. Sample image continua from painted to human (upper row) and CG to human faces (lower row) for Experiment 2. Morph percentages (% human) are shown below the images.

pseudorandomized order. The experiment was programmed and presented using E-Prime (Version 2.0; Psychology Software Tools).

Statistical analysis

Data reduction. Rating data for the human end point were replicated for both continua, and ratings were then pooled across participants and actors. All analyses were carried out in R (Version 3.5.2; R Core Team, 2016).

Model fitting. The following logistic function was fitted to the data using function *nlsLM* in R:

$$f(h) = \frac{c}{1 + e^{-a(h-b)}} - d \quad (1)$$

where h and f denote human-likeness and affinity, a denotes the steepness of the logistic curve, b denotes its middle point on the human-likeness axis, c denotes its range from minimum to maximum, and $-d$ denotes its minimum value (maximum value was $c - d$). Notably, h could denote either objective or subjective human-likeness depending on the analysis.

The following polynomial functions were fitted to the data using function *lm* in R:

$$f(h) = \sum_{i=0}^d (c_i h^i + d_i D h^i) \quad (2)$$

where d denotes the degree of the polynomial (1–4), c_i is the coefficient for polynomial regressor i , D is the dummy regressor for the continuum (0 for *painted* and 1 for *CG*), and d_i is the coefficient for dummy polynomial regressor i . For the final selected model, polynomial regressors h^i were orthogonalized using R function *poly*, which allowed us to test polynomial components independently of each other. Importantly, this procedure had no influence on model fits.

Model selection. Given that logistic and polynomial models were not nested, their model fits could not be compared using conventional methods. Instead, we adapted a Bayesian hypothesis testing approach for all model comparisons based on the guidelines of Masson (2011). Specifically, Bayesian Information Criterion (BIC) measures were first calculated for the null and alternative models M_0 and M_1 , and Bayes factor (BF) was then estimated as $\text{BF} \approx e^{(\Delta\text{BIC})/2}$. Probabilities for null and alternative models were calculated as $p(M_0|D) = \frac{\text{BF}}{\text{BF}+1}$ and $p(M_1|D) = 1 - p(M_0|D)$. Nominally, probabilities above .75 can be considered as positive evidence, above .95 as strong evidence, and above .99 as very strong evidence for a specific hypothesis (Masson, 2011; Raftery, 1999).

Results and Discussion

Subjective versus objective evaluations. Figure 6(a) illustrates relationships between subjective ratings and objective human-likeness levels for painted and CG continua. In H1, we predicted that a logistic function would fit the relationship between subjective ratings and objective human-likeness better than a cubic polynomial function. As can be seen in Table 1, the results provided very strong evidence in favor of the logistic model for painted

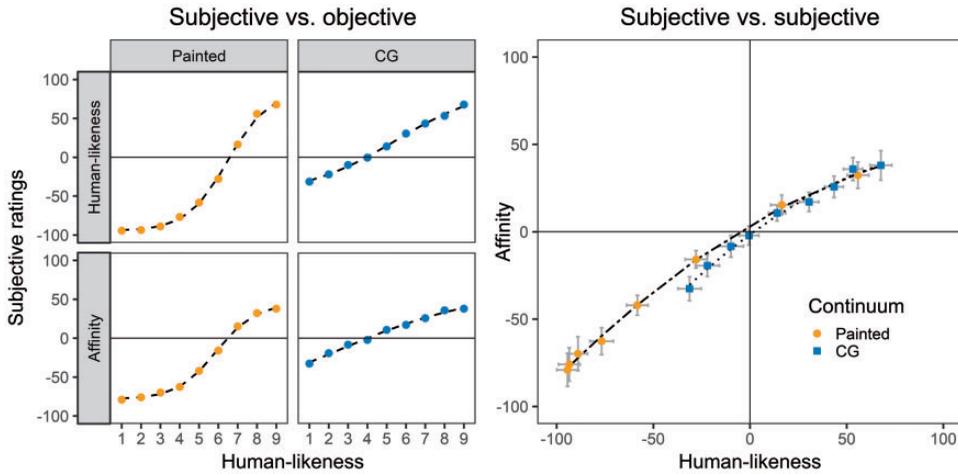


Figure 6. (a) Subjective ratings plotted against objective human-likeness by face-type continuum (painted and CG) and evaluation (human-likeness and affinity). Dashed lines illustrate fitted logistic curves. (b) Subjective affinity ratings plotted against subjective human-likeness ratings by continuum. Dashed lines illustrate fitted polynomial curves. Error bars illustrate within-subjects 95% confidence intervals (Morey, 2008). CG = computer-generated.

Table 1. Model Comparison Results for Subjective Ratings Versus Objective Human-Likeness Levels.

Continuum	Rating	Polynomial BIC	Logistic BIC	Δ BIC	BF	$p(\text{logistic} D)$	
Painted	HL	69.25	51.65	-17.61	1.50e-4	>.999	+++
Painted	Affinity	62.12	42.20	-19.92	4.73e-5	>.999	+++
CG	HL	41.27	43.11	1.84	2.51	.285	
CG	Affinity	44.49	45.33	0.84	1.53	.396	

Note. CG = computer-generated; HL = human-likeness; BIC = Bayesian Information Criterion; BF = Bayes factor; $p(\text{logistic}|D)$ = posterior probability for logistic over polynomial model given the data. +++ = very strong evidence for logistic model.

continuum but weak evidence for either model for CG continuum. This finding demonstrates that the relationship between subjectively perceived human-likeness and objective human-likeness manipulations is nonlinear, at least when the human-likeness scale is broad enough (e.g., from painted to human faces). The observed logistic pattern is consistent with the previously demonstrated categorical perception of human and nonhuman faces (Cheetham et al., 2011; Looser & Wheatley, 2010). Unlike these previous studies, however, we demonstrate that this logistic pattern also influences subjective affinity evaluations. This observation strengthens our suggestion that the UV curve should be tested against subjective rather than objective measures of human-likeness.

Uncanny slope versus UV. Figure 6(b) illustrates relationships between subjective affinity and subjective human-likeness for painted and CG continua. As can be seen in Table 2, a quadratic model provided clearly better fit to the data than linear model, but cubic and quartic models failed to improve model fit further. Consequently, the quadratic polynomial model was chosen as the best fitting model (adjusted $R^2 = .997$). Importantly, these findings

Table 2. Model Comparison Results for Subjective Affinity Versus Subjective Human-Likeness Ratings in Experiment 2.

M_0	M_1	BIC	Δ BIC	BF	$P(M_1 D)$	
Linear	Quadratic ^a	92.69	-28.34	7.00e-7	>.999	+++
Quadratic ^a	Cubic	95.75	3.06	4.63	.178	-
Quadratic ^a	Quartic	95.23	2.54	3.56	.219	-

Note. M_0 = reference model; M_1 = tested model; BIC = Bayesian Information Criterion estimate; BF = Bayes factor; $p(M_1|D)$ = posterior probability for tested over reference model given the data; +++ = very strong evidence for tested model; -- = positive evidence for reference model.

^aSelected model.

Table 3. Nonlinear Regression Results for the Selected Model for Subjective Affinity Versus Subjective Human-Likeness in Experiment 2.

Variable	Estimate	SE	T	p	η_p^2
Intercept	-8.71	0.85	-10.22	<.001	.897
Linear	167.27	2.84	58.99	<.001	.997
Quadratic	-24.54	3.40	-7.23	<.001	.813
Cont.	-10.19	2.55	-3.99	.002	.570
Cont. \times Linear	45.22	13.87	3.26	.007	.470
Cont. \times Quadratic	-11.51	9.72	-1.18	.259	.105

Note. Parameter estimates are based on orthogonalized polynomials for human-likeness ratings and are shown in arbitrary units. Cont. refers to a dummy variable for the continuum (0 for *painted* and 1 for *computed-generated*).

SE = standard error.

provided evidence against hypothesis H2b or the weak UV hypothesis, which predicted that a cubic relationship would provide the best fit to the data.

Results for the best fitting nonlinear regression model are shown in Table 3. Given that the linear component was significant, the results supported hypothesis H2a, which predicted an uncanny slope effect for the relationship between subjective affinity and subjective human-likeness. Unexpectedly, we also observed a statistically significant effect for the quadratic component. As can be seen in Figure 6(b), this effect is evident as a slight upward curvature of the more prominent slope effect. We interpret this to mean that decreases in human-likeness elicited decreasing affinity in a slightly accelerating manner. Finally, our hypothesis H3 predicted that CG faces would appear more eerie and that CG continuum would elicit a steeper uncanny slope than painted continuum. Visual inspection of Figure 6(b) suggests that CG faces elicited lower affinity than equally human-like points on the painted continuum. The statistically significant *Continuum* \times *Linear* effect in Table 3 consistently demonstrated that CG continuum elicited higher slope values than painted continuum.

Taken together, the present findings showed that subjective affinity for artificial faces decreases in a linear albeit slightly accelerating manner as their subjectively perceived human-likeness decreases. Consequently, the findings clearly supported an uncanny slope rather than a weak UV effect. However, we also found that the starting point of CG continuum elicited more negative affinity than equally human-like faces on the painted continuum, which supports the notion that CG faces appear particularly eerie (MacDorman & Chattopadhyay, 2016, 2017). These findings can be taken as tentative evidence for a weak UV effect in CG faces.

Experiment 3: Painted–CG–Human Continuum

Even though our second experiment provided initial evidence for a weak uncanny effect in CG faces, for a full demonstration, one would need to show that CG faces evoked a negative affinity peak with respect to neighboring human-likeness levels on both sides. The second experiment could not have provided such evidence given that the CG continuum did not extend beyond the CG faces on its left side. Therefore, in our third experiment, we investigated a new human-likeness continuum that extended from CG faces towards both painted and human faces. Specifically, faces were first transformed from painted to CG faces and then from CG to human faces (i.e., painted–CG–human continuum).

Our secondary aim was to fix potential methodological problems in the second experiment. First, we changed the bipolar human-likeness scale (from -100 to 100) to a unipolar scale (from 0 to 100). Our reasoning was that a bipolar scale might have artificially dichotomized participants' responses in the second experiment such that even slightly artificial faces always received below zero ratings. Second, whereas our affinity scale combined eeriness and pleasantness constructs (i.e., *unpleasant and creepy*), we now used a scale that focused only on eeriness. Third, we used a different computer-generation method for CG faces to increase the generalizability of our results. Fourth, human-likeness and affinity evaluations were made by two independent participant groups to avoid any carryover effects. Fifth, although less important, we sampled the subjective human-likeness space more evenly by using a separate pretest to select the final morph levels. We tested whether any of these changes would eliminate the quadratic component for subjective evaluations as observed in the second experiment. That is, we tested the hypothesis that the quadratic component would again be statistically significant for the relationship between subjective affinity and subjective human-likeness (H1).

The primary goal of this experiment was to provide evidence for a weak UV effect in CG faces. Hence, our main hypothesis was that CG faces would evoke more negative affinity than some of their neighboring levels on both sides of the human-likeness axis (H2). Given that we have already demonstrated an overall positive relationship between human-likeness and affinity (i.e., the uncanny slope effect), the critical question was whether CG faces would evoke more negative affinity than some of their less human-like neighbors.

Methods

Participants. Participants were 65 (53 females; $M = 19.9$ years, $SD = 1.9$ years) university students. Two additional participants who failed to pass diagnostic tests and had clearly deviant responses (see “Procedure” section in Experiment 3) were excluded from the study. All participants provided informed consent prior to the beginning of the experiment. Participants received course credit in compensation for their participation.

Stimuli. Stimuli were generated similarly as in Experiment 2 with the following changes. We used continua from painted to human faces (painted–human) and painted to human via CG faces (painted–CG–human), where the latter consisted of painted–CG and CG–human morph sequences. Unlike in Experiment 2, we now used FaceGen modeler to generate the CG faces. Morph percentages were selected such that human-likeness levels would cover the subjective human-likeness axis in roughly equidistant steps. To this end, nine novel pretest participants rated the human-likeness of initial painted–human and painted–CG–human continua consisting of 11 levels (10% morph step). Final morph percentages for 10 human-likeness levels, shown in Figure 7, were selected using linear interpolation for averaged ratings.

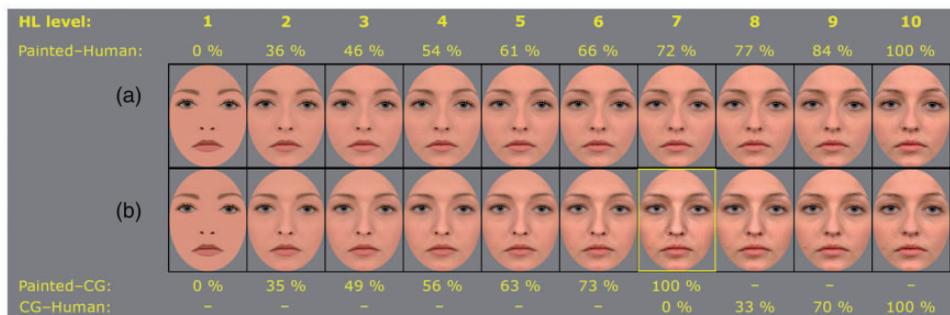


Figure 7. Sample images from (a) painted to human continuum and (b) painted to human via CG continuum for Experiment 3. CG face is emphasized with a yellow border. Morph percentages (e.g., % human) were selected such that subjective HL would increase in roughly equidistant steps. Morph percentages are shown above and below the respective continua. CG = computer-generated; HL = human-likeness.

Design. This study used a 2 (continuum: painted-human and painted-CG-human) \times 9 (human-likeness) \times 2 (rating: human-likeness and affinity) mixed design. Painted and human end points, which were common to both continua, were presented only once. Participants were assigned randomly to two groups rating either human-likeness ($N = 34$) or affinity ($N = 31$).

Procedure. This study was carried out as an online evaluation using Qualtrics (<http://www.qualtrics.com>). Only participants using a laptop or a desktop computer with a sufficiently large display (minimum 12") were accepted into the study. Given that the end points common to both continua were shown only once, participants saw and evaluated 180 stimuli (10 levels for painted-human + 8 levels for painted-CG-human \times 10 actors). Human-likeness was rated on a Visual Analogue Scale ranging from 0 (*not at all realistic*) to 100 (*completely realistic*) and affinity on a Visual Analogue Scale from -100 (*quite creepy*) to 100 (*quite nice*). We adopted *nice* as an opposite anchor for creepiness and used the adjective *quite* to encourage participants to use the whole rating scale. Six novel practice stimuli were evaluated before the actual study. We used two diagnostic tests to identify careless responders: We presented one instructed response item (*please respond exactly 42*) and asked participants to judge whether we should use their response data or not (cf. Meade & Craig, 2012). Participants failing to pass either of these tests (7 of the 67) were tagged for further inspection, and two participants whose responses showed close to zero variation were excluded from further analyses.

Statistical analysis

Model selection. Model selection for first- to fourth-degree polynomial models (with dummy regressors encoding the continua) was carried out in a similar manner to Experiment 2.

Comparisons between human-likeness levels. We first pooled affinity ratings across actors for each subject in the affinity group. Given that our data violated the assumption of sphericity for within-subjects analysis of variance, Mauchly's test: $\chi^2(44) = 365.44$, $p < .001$, we carried out a conservative nonparametric analysis using Friedman test. Pairwise comparisons between levels were carried out using the method of Eisinga, Heskens, Pelzer, and Te Grotenhuis (2017) as implemented in R function

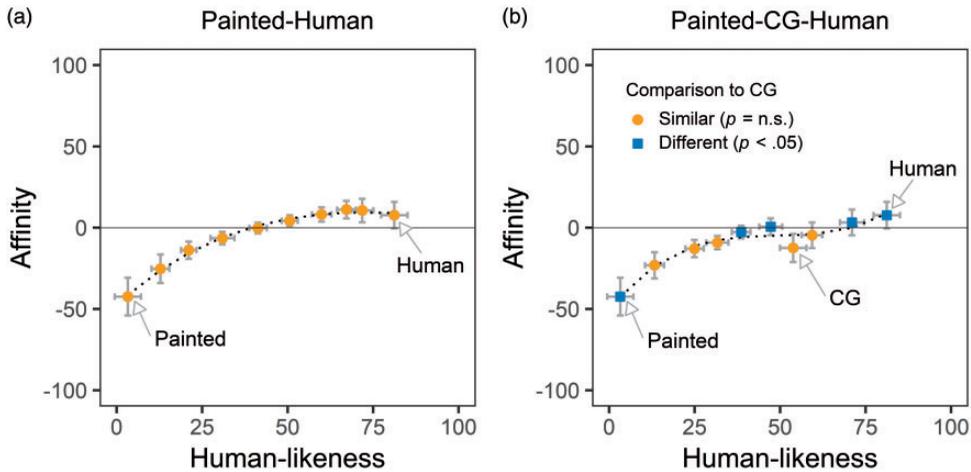


Figure 8. Plots between subjectively evaluated affinity and human-likeness in Experiment 3 for (a) painted to human and (b) painted to human via CG images. Dashed lines illustrate fitted third-degree polynomial curves, and error bars illustrate within-subjects 95% confidence intervals (Morey, 2008). In Panel (b), continuum levels similar to or different from CG faces (at $p < .05$, false discovery rate corrected) are emphasized using different colors and shapes. CG = computer-generated.

frdAllPairsExactTest, with false discovery rate correction ($q = .05$) applied for multiple comparisons.

Results and Discussion

Figure 8 illustrates the observed relationships between human-likeness and affinity. Visual inspection of Figure 8 suggests that painted–human continuum again elicited an uncanny slope effect with slight upward curvature (a quadratic effect), whereas the evaluation curve for painted–CG–human continuum instead resembled a weak UV (a cubic effect). These observations were confirmed by our nonlinear regression analyses. First, as can be seen in Table 4, cubic polynomial model provided the best fit to the data (adjusted $R^2 = .956$). As before, linear component was statistically significant in this model (Table 5). Consistently with H1, quadratic component was significant but did not differ between the continua. In other words, methodological changes (most notably, using a unipolar rather than a bipolar human-likeness scale) did not eliminate the slight curvature of the uncanny slope effect.

Second, as shown in Table 5, the cubic component was significantly stronger for painted–CG–Human than for painted–human continuum. As can be seen in Figure 8, this can be explained by a clearly deviant response to CG faces. To fully test H2, we next compared affinity ratings between CG faces and other human-likeness levels in painted–CG–human continuum. We observed a significant main effect of human-likeness level for affinity ratings both in painted–CG–human, Friedman test: $\chi^2(9) = 86.97$, $p < .001$, Kendall’s $W = .490$, and in painted–human, $\chi^2(9) = 115.08$, $p < .001$, $W = .573$, continuum. Confirming H2, pairwise comparisons showed that CG faces (Level 7) evoked significantly greater negative affinity than Levels 5 ($p = .019$) and 6 ($p = .008$) on its left side and Levels 9 ($p = .002$) and 10 ($p < .001$) on its right side (Figure 8(b)). Hence, CG faces fulfilled our criteria for eliciting a weak UV effect.

We also note that painted faces (Level 1) elicited significantly more negative affinity than any other level ($ps \leq .047$) in painted–human continuum and significantly more negative

Table 4. Model Comparison Results for Affinity Versus Human-Likeness Ratings in Experiment 3.

M_0	M_1	BIC	Δ BIC	BF	$p(M_1 D)$	
Linear	Quadratic	132.98	-14.51	1/1,415	.999	+++
Quadratic	Cubic ^a	122.03	-10.95	1/239	.996	+++
Cubic ^a	Quartic	125.63	3.59	6.03	.142	-

Note. M_0 = reference model; M_1 = tested model; BIC = Bayesian Information Criterion estimate; BF = Bayes factor; $p(M_1|D)$ = posterior probability for tested over reference model given the data. +++ = very strong evidence for tested model; - = positive evidence for reference model.

^aSelected model.

Table 5. Nonlinear Regression Results for the Selected Model for Affinity Versus Human-Likeness in Experiment 3.

Variable	Estimate	SE	T	p	η_p^2
Intercept	-4.66	1.07	-4.37	.001	.614
Linear	67.71	4.62	14.65	<.001	.947
Quadratic	-28.80	4.90	-5.88	<.001	.742
Cubic	3.35	4.74	0.71	.493	.040
Cont.	-4.96	1.51	-3.28	.007	.473
Cont. \times Linear	-13.00	6.75	-1.93	.078	.236
Cont. \times Quadratic	10.45	6.76	1.55	.148	.166
Cont. \times Cubic	15.48	6.74	2.30	.040	.306

Note. Parameter estimates are based on orthogonalized polynomials for human-likeness ratings and are shown in arbitrary units. Cont. refers to a dummy variable for the continuum (0 for *painted-human* and 1 for *painted-CG-human*).

SE = standard error.

affinity than any other level ($ps \leq .038$) except for Level 2 ($p = .163$) in *painted-CG-human* continuum. This shows that the least human-like faces rather than semirealistic faces (e.g., CG faces) elicited the most negative affinity. Although in Figure 8(a), it seems as if affinity would decrease slightly for human faces (Level 10), pairwise comparisons did not show any significant differences between Level 10 and any of the Levels 5 to 9 ($ps \geq .357$).

Taken together, this experiment provided evidence for a weak UV effect for CG faces. However, this effect was small in comparison to the uncanny slope effect, and *painted* faces received the most negative affinity. The quadratic component again showed that negative affinity increased in a slightly accelerating manner across decreasing human-likeness.

General Discussion

The present three studies show that overall, the subjective evaluation of artificial to human faces resembles an uncanny slope more than a UV. In other words, less human-like faces evoke more negative subjectively experienced affinity in a linear (but slightly accelerating) manner without evidence for a dip in subjective affinity. CG faces, which will be discussed later, are a possible exception to this pattern. This pattern, which we have referred to as the *uncanny slope*, is consistent with the bulk of previous empirical studies (e.g., Experiment 1 in Burleigh, Schoenherr, & Lacroix, 2013; Carter et al., 2013; Cheetham, Suter, & Jäncke, 2014; Looser & Wheatley, 2010; MacDorman, Green, Ho, & Koch, 2009; Rosenthal-von der Pütten & Krämer, 2014; Experiment 1 in Seyama & Nagayama, 2009). The present investigation significantly expands upon these studies, however. In particular, unlike

previous studies that have typically focused only on either CG faces (e.g., Cheetham et al., 2011; MacDorman & Chattopadhyay, 2016) or simplistic faces (e.g., Looser & Wheatley, 2010; Seyama & Nagayama, 2009), we explicitly investigated simplistic (painted) and CG faces in the same experiment. This allowed us to investigate the UV using a broad range of human-likeness on the one hand and to test whether CG faces are special with respect to other kinds of stimuli on the other hand. Methodologically, the present investigation was designed to avoid previous problems with naturalistic stimuli (e.g., the presence of various uncontrollable confound factors) and morphed images (e.g., artifacts caused by dissimilar source and target images).

We suggest that the uncanny slope effect observed in this and several previous studies may be explained simply by the much greater perceptual familiarity individuals have with human rather than artificial faces. Accordingly, we suggest that facial features in artificial faces are implicitly compared against those of typical human faces, and greater deviations from the human face prototype are then associated with greater negative affinity. This suggestion accommodates the fact that artificial faces are not a unified and naturally occurring category of objects—instead, artificial faces can appear artificial for an almost endless variety of reasons. This suggestion cannot explain the slight curvature of the subjective evaluation curve or the clearly deviant evaluation of CG faces, however.

The slight curvature of the uncanny slope was evident as a quadratic component in the evaluation curve between human-likeness and affinity in two different studies that used different methods (Figures 6(b) and 8(a)). To the best of our knowledge, a similar finding has not been observed in previous empirical studies. We suggest that the quadratic component could be explained by a slight preference for intermediate human-likeness. This suggestion is consistent with the original UV hypothesis (Mori, 1970/2012), which predicted that a positive affinity peak occurs for some of the moderately human-like artificial entities. It is, however, important to note that contrary to our results, Mori predicted a nonmonotonic function in which affinity first increases and then decreases after this initial peak. In contrast, we observed a monotonic change in which the first derivative varied but affinity always increased across increasing human-likeness throughout the whole human-likeness axis. Importantly, the least human-like faces elicited the most negative affinity, and completely human faces elicited the most positive affinity. Our observed result pattern, despite its nonlinearity, is hence consistent with our characterization of the uncanny slope effect.

CG faces, which evoked more negative responses than their neighbors on both sides (Figure 8(b)), provided a clear exception to the uncanny slope pattern. This result could possibly be explained by the previous observation that realism inconsistency between individual facial features evokes negative affinity (MacDorman & Chattopadhyay, 2016; Seyama & Nagayama, 2009). In our painted continua, all faces were morphed evenly. In contrast, CG faces were likely to contain some facial features that were less realistic than others, as is typically the case for CG faces (e.g., MacDorman & Chattopadhyay, 2017). To the best of our knowledge, our findings demonstrate for the first time that CG faces can actually elicit more negative affinity than other less realistic stimuli. This adds significantly to previous studies that have investigated either CG (e.g., MacDorman & Chattopadhyay, 2016) or other less realistic stimuli (e.g., Seyama & Nagayama, 2009) but not explicitly compared them against each other. We also provide a theoretical contribution by considering the distinction between weak and strong forms of the UV hypothesis and by suggesting that the present findings support only a weak form of the UV effect for CG faces. In other words, even though CG faces are slightly uncanny in comparison to their neighbors, this effect is weak when compared with the more prominent uncanny slope pattern.

Previous studies have already demonstrated that continua between artificial and human faces are perceived categorically, and therefore they are related to subjective human-likeness judgments via a logistic rather than a linear function (Cheetham et al., 2011; Looser & Wheatley, 2010; MacDorman & Chattopadhyay, 2016). We extend these findings by showing that this logistic pattern also affects subjective affinity ratings (Figure 6(a)). This means that other superimposed effects are difficult or impossible to segregate from this logistic pattern if subjective affinity ratings are compared against objectively manipulated human-likeness (e.g., as in Thompson et al., 2011). On the other hand, the same objective human-likeness scale should not be used as a common metric for quantitatively different human-likeness continua. For example, the same *percentage human* scale is not necessarily meaningful for realism-consistent and realism-inconsistent image morphs (e.g., MacDorman & Chattopadhyay, 2016), given that these continua may evoke different changes in subjectively perceived human-likeness. Methodologically, these observations can be taken to suggest that future UV studies should contrast subjective affinity against subjective rather than objective measures of human-likeness, unless they can show that the latter two are roughly identical and that they employ only one type of stimulus continuum.

We discuss some limitations of the present study and make suggestions for future research. We investigated static images of faces, even though dynamic stimuli might have evoked stronger effects. In particular, Mori (1970/2012) predicted in his original UV essay that movement would amplify the UV curve. We would like to defend the use of a static modality, however, given that in our view the UV hypothesis should first be confirmed with static stimuli before considering any other modulatory effects. It could also be argued that because we investigated passively observed face stimuli, our findings were not truly representative of the UV. It should be noted, however, that Mori's original essay also focused on passive observation. Face stimuli are also particularly salient for the UV, given that faces are innately social and the human neural system is highly specialized in processing facial information (e.g., Haxby, Hoffman, & Gobbini, 2000). A related limitation is that we focused on rigorously controlled experimental stimuli, which may limit the generalizability of our results to naturalistic contexts (e.g., encounters with physical robots in socially interactive contexts). We opted to use controlled stimuli to avoid confound effects that are in our view unavoidable with naturalistic stimuli. The issue of using well-controlled versus ecologically valid stimuli still remains open to debate, however.

Even though we were able to tap into a broad range of human-likeness by incorporating painted faces into our experimental designs, the painted faces were still recognizably human-like. It is conceivable that at least the first positive affinity peak of the UV would be more pronounced if the human-likeness continua began from clearly nonhuman stimuli such as random shapes. With this kind of continuum, the mere recognition that some of the stimuli begin to resemble humans might evoke feelings of familiarity and positive affinity in human observers. This hypothesis could be pursued in future studies; however, generating well-controlled stimulus continua from nonhuman to human remains a considerable methodological challenge.

Importantly, we demonstrated a similar weak UV pattern for morph sequences beginning from two different kinds of CG faces (FaceGen and MakeHuman) in Experiments 2 and 3, which increased our confidence in that this result was not simply related to a particular CG face generation method. In contrast, we only employed one type of a painted face, which reduces the generalizability of our results. It is, however, worth noting that our painted-human continuum replicated the uncanny slope pattern already demonstrated in several previous studies (cf. Kättsyri et al., 2015). Furthermore, our painted-human and CG-human continua evoked similar affinity curves with the exception that painted faces were actually

considered more familiar than equivalent faces on the CG–human continuum. This somewhat alleviates the concern that the (curved) uncanny slope could have resulted from peculiarities with the painted faces. Future studies with other kinds of simplistic faces should be carried out to replicate the present findings, however.

A further limitation of the present studies is that we did not consider the role of aesthetics and design intent on our findings. The idea that purposefully aesthetic designs could be used to overrule the UV effect is not a new one (cf. Hanson, 2005). Previous evidence also suggests that the aesthetic design potential of artificial characters varies depending on their human-likeness. In particular, in line with traditional animation design principles (Lasseter, 1987), simplification and exaggeration can be used to create more appealing virtual characters and robots but only when they are sufficiently *nonhuman-like*. In human-like characters, exaggeration seems to have the opposite effect of making them appear eerie (e.g., Green, MacDorman, Ho, & Vasudevan, 2008; Mäkäräinen, Kätsyri, & Takala, 2014). Aesthetic design potential would explain why in some recent studies, contrary to our findings with painted faces, cartoonish faces with neonatal features evoked more positive affinity than realistic CG faces (Schindler et al., 2017; Zell et al., 2015). One possibility, which could be investigated in future studies, is that the UV occurs because the principles of simplification and exaggeration are only applicable to at most moderately human-like artificial characters.

The present findings are of methodological significance for behavioral and neuroscientists, in particular because we investigated CG stimuli generated with methods that are easily accessible and widely used in empirical research (e.g., Krumhuber et al., 2012). We clearly demonstrated that, in comparison to human faces, all artificial faces evoke negative affinity to some extent, and that this negative affinity increases in an accelerating manner across decreasing human-likeness (i.e., curved uncanny slope effect). This suggests that real human faces are always better research stimuli than CG faces, in particular because the latter evoke a weak form of the UV (Figure 8(b)). At the same time, we did not observe evidence for a strong UV in which CG faces would have evoked more negative affinity than any other stimuli. Together with the uncanny slope, this observation suggests that trying to purposefully avoid too high a level of realism would actually be counterproductive. Hence, when CG faces are used in lieu of human face as experimental stimuli, they should be made as realistic as possible, as long as their individual features do not represent highly dissimilar realism levels. Future studies are still required to answer the question of whether greater realism inconsistency is unavoidable in increasingly realistic CG faces. Highly schematic faces could also be advantageous in some research settings to isolate the features of interest for the researcher (cf. de Gelder, Kätsyri, & de Borst, 2018).

The present findings lead to similar recommendations for practical applications. Previously, we tentatively suggested that simplification and exaggeration can be used to make nonhuman-like virtual characters and robots more appealing. This is not a viable approach for practical applications for which realistic virtual characters or robots are desirable (e.g., conversational interfaces or digital games with virtual humans). The present uncanny slope pattern suggests that all virtual characters that are recognizable as artificial always evoke some degree of negative affinity. At the same time, this pattern clearly implies that trying to avoid realism aggravates rather than alleviates this problem. Returning to our hypothetical opening example, trying to avoid the UV in the film *Life of Pi* by reducing the realism of the computer-animated tiger scenes would likely evoke higher levels of negative affinity in viewers.

To conclude, the present three studies supported an uncanny slope effect rather than a UV hypothesis for the subjective evaluation of virtual faces: the less realistic the virtual faces

appear the more negative they become. This evaluation curve was slightly curved, which possibly resulted from a preference for intermediate human-likeness levels. As a possible exception to the uncanny slope pattern, CG faces evoked a weak UV effect in which they elicited more negative affinity than equally human-like faces on a continuum from painted to human faces. This effect was much weaker in comparison to the global uncanny slope pattern, however. In particular, the most artificial faces always evoked the most negative affinity, and the most realistic faces always evoked the most positive affinity. Contrary to the strong variant of the UV hypothesis, these findings tentatively encourage the development of increasingly realistic virtual characters and robots for research as well as for practical applications.

Authors' Note

Experimental setup, data preprocessing, and data analysis files are available in <https://osf.io/5zccq9/> with the exception of stimulus images due to Radboud face database copyright transfer restrictions.

Acknowledgements

The authors would like to thank Richard Benning, BICT, Maastricht University, Instrumentation Department, for his help in producing the computer-generated faces used in the present experiment.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 703493 *NeuroBukimi* and from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) grant agreement No. 295673 *EMOBODIES*.

ORCID iD

Jari Kätsyri  <https://orcid.org/0000-0002-7003-1238>

Tapio Takala  <https://orcid.org/0000-0002-7704-5800>

Supplemental Material

Supplementary material for this article is available online.

Note

1. Cronbach's α s were computed separately for each condition, standardized using Fisher Z-transformation, averaged, and inverse-transformed.

References

- Almaraz, S. M. (2017). *Uncanny processing: Mismatches between processing style and featural cues to humanity contribute to uncanny valley effects* (Master's thesis). Miami University, Oxford, OH.
- Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in Human Behavior*, 52, 331–337. doi:10.1016/j.chb.2015.06.018

- 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*, 759–771. doi:10.1016/j.chb.2012.11.021
- Butler, M., & Joschko, L. (2009). Final fantasy or the Incredibles. *Animation Studies*, *4*, 55. doi:10.1201/9780429450716-1
- Carr, E. W., Hofree, G., Sheldon, K., Saygin, A. P., & Winkielman, P. (2017). Is that a human? Categorization (dis)fluency drives evaluations of agents ambiguous on human-likeness. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 651–666. doi:10.1037/xhp0000304
- Carter, E. J., Mahler, M., & Hodgins, J. K. (2013, August 22–23). Unpleasantness of animated characters corresponds to increased viewer attention to faces. In Proceedings of the ACM Symposium on Applied Perception (pp. 35–40). New York, NY: ACM. doi:10.1145/2492494.2502059
- Chattopadhyay, D., & MacDorman, K. F. (2016). Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of Vision*, *16*, 7. doi:10.1167/16.11.7
- Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the “uncanny valley hypothesis”: Behavioral and functional MRI findings. *Frontiers in Human Neuroscience*, *5*, 126. doi:10.3389/fnhum.2011.00126
- Cheetham, M., Suter, P., & Jäncke, L. (2014). Perceptual discrimination difficulty and familiarity in the uncanny valley: More like a “happy valley.” *Frontiers in Psychology*, *5*, 1219. doi:10.3389/fpsyg.2014.01219
- de Gelder, B., Kätsyri, J., & de Borst, A. W. (2018). Virtual reality and the new psychophysics. *British Journal of Psychology*, *109*, 421–426. doi:10.1111/bjop.12308
- Deska, J. C., Almaraz, S. M., & Hugenberg, K. (2017). Of mannequins and men: Ascriptions of mind in faces are bounded by perceptual and processing similarities to human faces. *Social Psychological and Personality Science*, *8*, 183–190. doi:10.1177/1948550616671404
- Eisinga, R., Heskes, T., Pelzer, B., & Te Grotenhuis, M. (2017). Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers. *BMC Bioinformatics*, *18*, 68. doi:10.1186/s12859-017-1486-2
- Fan, S., Wang, R., Ng, T.-T., Tan, C. Y.-C., Herberg, J. S., & Koenig, B. L. (2014). Human perception of visual realism for photo and computer-generated face images. *ACM Transactions on Applied Perception*, *11*, 1–21. doi:10.1145/2620030
- Green, R. D., MacDorman, K. F., Ho, C.-C., & Vasudevan, S. (2008). Sensitivity to the proportions of faces that vary in human likeness. *Computers in Human Behavior*, *24*, 2456–2474. doi:10.1016/j.chb.2008.02.019
- Hanson, D. (2005, December 4). *Expanding the aesthetic possibilities for humanoid robots*. Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots, Tsukuba.
- Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*, 223–233. doi:10.1016/S1364-6613(00)01482-0
- Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, *26*, 1508–1518. doi:10.1016/j.chb.2010.05.015
- Ho, C.-C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. *International Journal of Social Robotics*, *9*, 129–139. doi:10.1007/s12369-016-0380-9
- Hugenberg, K., Young, S., Rydell, R. J., Almaraz, S., Stanko, K. A., See, P. E., & Wilson, J. P. (2016). The face of humanity: Configural face processing influences ascriptions of humanness. *Social Psychological and Personality Science*, *7*, 167–175. doi:10.1177/1948550615609734
- Kageki, N. (2012). An uncanny mind. *IEEE Robotics Automation Magazine*, *19*, 112–108. doi:10.1109/MRA.2012.2192819
- Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*, 390. doi:10.3389/fpsyg.2015.00390

- Kätsyri, J., Mäkäräinen, M., & Takala, T. (2017). Testing the ‘uncanny valley’ hypothesis in computer-animated film characters: An empirical evaluation of natural film stimuli. *International Journal of Human-Computer Studies*, *97*, 149–161. doi:10.1016/j.ijhcs.2016.09.010
- Kobayashi, M., Otsuka, Y., Kanazawa, S., Yamaguchi, M. K., & Kakigi, R. (2012). Size-invariant representation of face in infant brain: An fNIRS-adaptation study. *NeuroReport*, *23*, 984–988. doi:10.1097/WNR.0b013e32835a4b86
- Krumhuber, E. G., Tamarit, L., Roesch, E. B., & Scherer, K. R. (2012). FACSGen 2.0 animation software: Generating three-dimensional FACS-valid facial expressions for emotion research. *Emotion*, *12*, 351–363. doi:10.1037/a0026632
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, *24*, 1377–1388. doi:10.1080/02699930903485076
- Lasseter, J. (1987). Principles of traditional animation applied to 3D computer animation. *ACM Siggraph Computer Graphics*, *21*, 35–44. doi:10.1145/37401.37407
- Lee, A. (2012). *Life of Pi* [Motion picture]. Los Angeles, CA: 20th Century Fox.
- Lischetzke, T., Izydorczyk, D., Hüller, C., & Appel, M. (2017). The topography of the uncanny valley and individuals’ need for structure: A nonlinear mixed effects analysis. *Journal of Research in Personality*, *68*, 96–113. doi:10.1016/j.jrp.2017.02.001
- Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science*, *21*, 1854–1862. doi:10.1177/0956797610388044
- 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. doi:10.1016/j.cognition.2015.09.019
- MacDorman, K. F., & Chattopadhyay, D. (2017). Categorization-based stranger avoidance does not explain the uncanny valley effect. *Cognition*, *161*, 132–135. doi:10.1016/j.cognition.2017.01.009
- 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*, 695–710. doi:10.1016/j.chb.2008.12.026
- MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, *7*, 297–337. doi:10.1075/is.7.3.03mac
- Mäkäräinen, M., Kätsyri, J., & Takala, T. (2014). Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley? *Cognitive Computation*, *6*, 708–721. doi:10.1007/s12559-014-9273-0
- Mangan, B. (2015). The uncanny valley as fringe experience. *Interaction Studies*, *16*, 193–199. doi:10.1075/is.16.2.05man
- Marschner, L., Pannasch, S., Schulz, J., & Graupner, S.-T. (2015). Social communication with virtual agents: The effects of body and gaze direction on attention and emotional responding in human observers. *International Journal of Psychophysiology*, *97*, 85–92. doi:10.1016/j.ijpsycho.2015.05.007
- Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690. doi:10.3758/s13428-010-0049-5
- 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. doi:10.1016/j.cognition.2015.09.008
- Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*, 255–260. doi:10.1016/S1364-6613(02)01903-4
- 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*, *31*, 1–11. doi:10.1145/2185520.2185587
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455. doi:10.1037/a0028085
- Moore, R. K. (2012). A Bayesian explanation of the ‘uncanny valley’ effect and related psychological phenomena. *Scientific Reports*, *2*, 864. doi:10.1038/srep00864

- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4, 61–64. doi:10.20982/tqmp.04.2.p061
- Mori, M. (1970/2012). The uncanny valley (K. F. MacDorman & N. Kageki, Trans.). *IEEE Robotics & Automation Magazine*, 19, 98–100. doi:10.1109/MRA.2012.2192811
- Piwiek, 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. doi:10.1016/j.cognition.2013.11.001
- Poliakoff, E., Beach, N., Best, R., Howard, T., & Gowen, E. (2013). Can looking at a hand make your skin crawl? Peering into the uncanny valley for hands. *Perception*, 42, 998–1000. doi:10.1068/p7569
- Raftery, A. E. (1999). Bayes factors and BIC: Comment on “A critique of the Bayesian information criterion for model selection.” *Sociological Methods & Research*, 27, 411–427. doi:10.1177/0049124199027003005
- Railo, H., Karhu, V.-M., Mast, J., Pesonen, H., & Koivisto, M. (2016). Rapid and accurate processing of multiple objects in briefly presented scenes. *Journal of Vision*, 16, 8. doi:10.1167/16.3.8
- R Core Team. (2016). *R: A language and environment for statistical computing*. Retrieved from <https://www.R-project.org/>
- Rosenthal-von der Pütten, A. M., & Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. *Computers in Human Behavior*, 36, 422–439. doi:10.1016/j.chb.2014.03.066
- Sasaki, K., Ihaya, K., & Yamada, Y. (2017). Avoidance of novelty contributes to the uncanny valley. *Frontiers in Psychology*, 8, 1792. doi:10.3389/fpsyg.2017.01792
- Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, 44, 718–730. doi:10.1016/j.neuropsychologia.2005.07.017
- Schindler, S., Zell, E., Botsch, M., & Kissler, J. (2017). Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Scientific Reports*, 7, 45003. doi:10.1038/srep45003
- Seyama, J., & Nagayama, R. S. (2009). Probing the uncanny valley with the eye size aftereffect. *Presence: Teleoperators and Virtual Environments*, 18, 321–339. doi:10.1162/pres.18.5.321
- Thompson, J. C., Traflet, J. G., & McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception*, 40, 695–704. doi:10.1068/p6900
- Wang, S., Lilienfeld, S. O., & Roachat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology*, 19, 393–407. doi:10.1037/gpr0000056
- Wang, S., & Roachat, P. (2017). Human perception of animacy in light of the uncanny valley phenomenon. *Perception*, 46, 1386–1411. doi:10.1177/0301006617722742
- Weis, P. P., & Wiese, E. (2017). Cognitive conflict as possible origin of the uncanny valley. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61, 1599–1603. doi:10.1177/1541931213601763
- Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the “uncanny valley” phenomenon. *Japanese Psychological Research*, 55, 20–32. doi:10.1111/j.1468-5884.2012.00538.x
- Zell, E., Aliaga, C., Jarabo, A., Zibrek, K., Gutierrez, D., McDonnell, R., & Botsch, M. (2015). To stylize or not to stylize? The effect of shape and material stylization on the perception of computer-generated faces. *ACM Transactions on Graphics*, 34, 1–12. doi:10.1145/2816795.2818126