Heliyon

Received:
24 July 2018
Revised:
11 October 2018
Accepted:
9 November 2018

Cite as: Jussi Palomäki, Anton Kunnari, Marianna Drosinou, Mika Koverola, Noora Lehtonen, Juho Halonen, Marko Repo, Michael Laakasuo. Evaluating the replicability of the uncanny valley effect. Heliyon 4 (2018) e00939. doi: 10.1016/j.heliyon.2018.e00939

Evaluating the replicability of the uncanny valley effect

Jussi Palomäki [a],[*], Anton Kunnari [b], Marianna Drosinou [b], Mika Koverola [a], Noora Lehtonen [a], Juho Halonen [a], Marko Repo [a], Michael Laakasuo [a],[1]

[a] *University of Helsinki, Faculty of Arts, Department of Digital Humanities, Cognitive Science, Siltavuorenpenger 1 A, 00012, Helsingin yliopisto, Finland*

[b] *University of Helsinki, Faculty of Medicine, Department of Psychology, Helsinki, Finland*

[*] Corresponding author.

E-mail address: jussi.palomaki@helsinki.fi (J. Palomäki).

[1] Dr. Laakasuo is the project PI.

Abstract

The uncanny valley (UV) effect refers to an eerie feeling of unfamiliarity people get while observing or interacting with robots that resemble humans almost but not quite perfectly. The effect is not well understood, and it is also unclear how well results from previous research on the UV can be replicated. In six studies, both in the laboratory and online (N = 1343), we attempted to replicate the UV effect with various stimuli used in previous research. In Studies 1 and 2 we failed to replicate the UV effect with CGI stimuli created using the so-called morphing technique (a robot image morphed into a human image, resulting in a supposedly creepy robot-human image). In Studies 3a and 3b we found a prominent UV effect using pre-evaluated, non-morphed and photorealistic robot pictures. Finally, in exploratory Studies 4a and 4b we found the UV effect using morphed and photorealistic human and robot pictures. Our results suggest that the UV effect is more robust when elicited by pre-validated or *prima facie* uncanny robot pictures than by non-photorealistic images generated using the morphing technique. We argue that photorealistic pictures are more suitable than less realistic CGI pictures as stimuli for research attempting to elicit the UV effect – however, our results do not invalidate any previous research on the UV effect using morphing techniques, but point to their domain of applicability and context sensitivity.

Keyword: Psychology

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

1. Introduction

The uncanny valley (UV) effect refers to an eerie feeling of unfamiliarity people get while observing or interacting with robots that resemble humans almost but not quite perfectly (Mori, 1970). To our knowledge there is generally a lack of pre-validated stimulus materials known to reliably, robustly and repeatedly elicit the UV effect, such as a common repository of "uncanny" robot pictures (or other objects) — or specific guidelines to create such stimuli. The primary aim in our present six studies was to find stimulus pictures that reliably and consistently elicit the UV effect without the stimuli being sensitive to contextual issues (e.g., which stimuli people have seen previously). This would be a valuable asset for researchers studying the phenomenon (see also Rosenthal-von der Pütten and Krämer, 2014).

The UV effect has been studied extensively for decades but there is no consensus on what gives rise to it, and several explanations have been proposed. Some explanations drawing from sexual selection and pathogen avoidance models propose that uncanny robots convey subliminal signals of "unfit" physical deformities (MacDorman et al., 2009; Mitchell et al., 2011); others draw from terror management theory and suggest that uncanny robots are salient reminders of our own mortality (MacDorman, 2005).

A promising hypothesis on the UV effect postulates that the effect is not specific to human-likeness *per se*, but is instead related to a difficulty of categorizing an almost familiar object when there are multiple competing "interpretations" of what that object could be. Such stimulus-category competition is ostensibly cognitively demanding and thus elicits negative affect (e.g., Ferrey et al., 2015). Accordingly, the UV effect could arise in any event in which some object appears to be lingering between two different object categories.

If the stimulus-category hypothesis is correct, producing images of uncanny objects should be possible by gradually morphing one image category into another (such as, but not limited to, robots gradually morphing into humans). This morphing technique has been tested in a number of studies, and the results have been promising (Ferrey et al., 2015; Matsuda et al., 2012; Yamada et al., 2012, 2013). For example, Yamada and colleagues (2013) found a UV effect for cartoon face images morphed with real face images, but not for two morphed real face images. Matsuda and colleagues (2012) showed that children dislike images of their mothers' faces morphed with strangers' faces but not two morphed strangers' faces. In another recent study, Ferrey and colleagues (2015) found support for the stimulus-category hypothesis by measuring the likability ratings for different gradual line drawings depicting non-human animal morphs (e.g., a duck gradually morphing into a rabbit over a series

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

of images). In these gradually morphing image series the mid point was always maximally bistable (and thus difficult to categorize, such as a “rabbit-giraffe”) and also the least likable. Ferrey and colleagues (ibid.) observed the same effect for human-robot- and various human-animal morphs.

However, a recent review by Kätsyri and colleagues (2015) highlighted a number of potential caveats in morphing techniques, such as the need to abide by specific and consistent guidelines in creating the morphed images. In earlier research, Hanson (2006) demonstrated that whether or not the UV effect is elicited by viewing morphed images depends on the aesthetic design of the images. Moreover, Seyama and Nagayama (2007) found that using the morphing technique for human faces supported the UV effect only when the faces had abnormal features, such as bizarre eyes. In a similar vein, Green and colleagues (2008) found that human-android morphs exaggerate or attenuate the UV effect depending on specific subtle changes in the image facial proportions.

Cheetham and colleagues (Cheetham et al., 2011, 2014) have shed light on how perception discrimination across morphed images influences the perceived affective valence of the images, and, consequently, the UV effect elicited by those images. For example, on a continuum of morphed images (with a constant “morph-distance” between each image), the maximally ambiguous images near the midpoint of the continuum were, in fact, the easiest to discriminate from nearby images (indicating a “happy valley” effect; Cheetham et al., 2014). These findings challenge the assumption that greater perceptual discrimination difficulty evokes negative affect and the UV effect. Moreover, a recent study by MacDorman and Chattopadhyay (2016) found that the UV effect is driven more by consistency in human realism than by perceived category uncertainty; they found that the eeriest and coldest rated human faces were also the least ambiguous, but entailed low realism in the eyes, eyelashes and mouth.

To our knowledge, the stimulus materials created using the morphing technique in one study have not been used to elicit the UV effect in different populations. In other words, the replicability and generalizability of the UV effect for these particular stimuli is unclear. Given the current and rampant replication crisis in psychological sciences (see Open Science Collaboration, 2015), it is important that independent researchers at least conceptually replicate existing studies.

However, we found this to be surprisingly difficult in the first two studies reported in this paper. Specifically, we present attempted conceptual replications of the findings of Ferrey and colleagues (2015) and on the stimulus-category hypothesis in a Finnish population for the human-robot -morphs, which are the stimuli of interest for most UV related research. Our choice to focus on the robot morphs was also motivated by the rapidly increasing number of robots across societies working on a plethora of tasks from healthcare to law enforcement. Moreover, the International Federation of Robotics has estimated that by 2019 more than 42 million robots have been sold for personal use; meaning, they are quickly becoming an unavoidable part of our social ecosystem (International Federation of Robotics, 2016).

We sought to evaluate the absolute value of the UV effect elicited by the stimuli created by Ferrey and colleagues (2015); but our aims were not to test the stimulus category hypothesis *per se*. Originally, we sought to use these human-robot -morphs to elicit the UV effect and observe its influence on subsequent decision-making tasks. However, our findings (unpublished manuscript) on the influence of the UV effect on decision-making tasks were inconclusive, and we reasoned this was probably because the stimuli did not elicit the UV effect in our study population. Since our original aim was not to fully replicate the results of Ferrey and colleagues (2015), our current studies should be viewed as conceptual replications; that is, attempts to replicate specific results with slightly differing methodology, while maintaining conceptual similarity with the original research. We were specifically interested in the UV effect in the context of robot images, since our initial goal was not to investigate the UV hypothesis per se, but to apply validated UV images in another context.

In studies 3a-b and 4a-b we also applied findings from other research (Rosenthal-von der Pütten and Krämer, 2014), where robots of varying levels of uncanniness were categorized based on pre-evaluation instead of being generated in a “theory-driven” manner – like using the morphing technique. Using these stimuli we were able to elicit the UV effect in our participants. Our results suggest that the UV effect is generally much more robustly elicited by pre-validated or *prima facie* selected robot pictures than by pictures generated using the morphing technique. It seems, in our case, that face validity and careful reflective selection of stimulus materials is more effective than using the morphing technique to elicit the UV effect across different contexts; we also argue that photorealistic pictures of both robot and human agents are most suitable for this purpose, especially as anchoring points for the continuum.

2. Materials & methods

2.1. Ethics statement

All local laws regarding ethics for social science research were followed in full in all studies. All participation was fully voluntary and participants were informed about their right to opt out at any point without penalties; informed consent was obtained from all participants in all studies. Our study protocols have been previously reviewed and approved by the University of Helsinki Ethical Review Board in Humanities and Social and Behavioural Sciences (project “Humans and Technology”, statement numbers 9/2017 and 48/2017).

2.2. Materials & methods of study 1

In study 1 we sought to conceptually replicate (i.e. with similar but not identical stimulus materials) the findings of Ferrey and colleagues (2015). More specifically, we attempted to elicit the UV effect with pictures generated with a morphing technique from non-photorealistic computer generated images (CGIs).

We set up a laboratory in a public library in Espoo, Finland, and recruited 221 participants (107 males, 114 females). Our laboratory conformed to usual standards of social psychological research: the environment was calm, quiet, and private, and we used a computer program that could not be aborted by the participant. Of the participants, about 75% had some level of university education while 55% described their income level as average or less (compared with what they estimated was the population average). The mean age of the participants was 38.7 years ($SD = 16.8$; range $= 18-80$). Our participants were thus library users in the second biggest city of Finland and much more representative of the general population than the average studies conducted in experimental psychology. Participants were paid 2.5 euros for participation.

Participants first gave informed consent and were then escorted into cubicles where they put on headphones playing low volume pink noise to filter out potential auditory distractions. In the cubicles, participants operated a 15.6″ laptop computer with a mouse. The study had a between-subjects experimental design aimed at evaluating how the perceived uncanniness of an agent influences subsequent decision-making. However, in the current study we focus only on the perceived uncanniness of the agents. Participants were randomly assigned to either evaluate a computer rendered 1) CGI human figure, 2) human-robot morph (closer to human), 3) robot-human morph (closer to robot), and 4) robot. See Fig. 1 for pictures of the agents, which were adopted directly from Ferrey and colleagues (2015). Thus, in this study we employed four images. However, conceptually, the minimum number of images needed for eliciting the UV effect should be three images: two at the "hilltops" (i.e. fully human and fully robot) and one at the bottom of the valley (uncanny image). This enables testing for a statistical quadratic effect (human + robot vs. uncanny image).

The picture was positioned in the middle of the screen against a black background with a Likert scale shown directly below it. Participants gave their responses by clicking the appropriate number on the Likert scale with a mouse. Each participant evaluated only one agent on the following characteristics, on a Likert 1 (Not at all) to 7 (Very) scale: *Is this agent 1) disgusting, 2) pleasant, 3) scary, 4) trustworthy, 5) repulsive, 6) creepy, 7) approachable, 8) friendly, and 9) nice.*

2.3. Materials & methods of study 2

In Study 2, we aimed to replicate the findings of Study 1 (the inability to find prominent UV effects). Since one of the reasons for the lack of effects in Study 1 could

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

have been the human CGI face not appearing realistic enough, we also added a pre-validated non-CGI human face image obtained from the Radboud database as a stimulus (Langner et al., 2010).

Our participant pool, compensation, and data collection procedure were the same as in Study 1, but repeat participation was deterred. In Study 2 we recruited 172 participants (76 males, 96 females). Of the participants, 50% had some level of university education while 54% described their income level as average or less (compared to what they believed was the population average). The mean age of the participants was 36.9 years ($SD = 16.6$; range = 18−78).

Participants were randomly assigned to either evaluate a 1) human figure obtained from the Radboud database, ("Human") 2) human CGI character ("Human-CGI"; identical to that in Study 1); 3) human-robot morph, 4) robot-human morph, and 5) robot. See Fig. 1 for figures of the agents. Each participant evaluated only one agent on the following characteristics, on a Likert 1 (Not at all) to 7 (Very) scale: *Is this agent 1) disgusting, 2) pleasant, 3) scary, 4) trustworthy, 5) repulsive, 6) creepy, 7) approachable, 8) friendly, and 9) nice.* Additionally, participants evaluated the agents on how *human/humane* they appeared. This was done to see whether there was a meaningful difference in perceived "humanness" between the CGI and actual human agents.

2.4. Materials & methods of study 3a

Since we could not consistently elicit the UV effect using the morphed images created by Ferrey and colleagues (2015), we searched for studies with pre-validated non-morphed robot pictures of varying levels of uncanniness. We found one such study by Rosenthal-von der Pütten and Krämer (2014), who had categorized an array of robot pictures into clusters based on their perceived uncanniness (among other attributes). In Study 3a, our aim was to find out whether we could elicit the UV effect with some of these pre-selected robot pictures.

Again, our participant pool, compensation, and data collection procedure were the same as in Studies 1 and 2. In Study 3a we recruited 125 participants (44 males, 81 females). About 56% had some level of university education while 84% described their income level as average or less. The mean age of the participants was 29.9 years ($SD = 9.6$; range = 18−71).

In Study 3a we obtained our stimulus materials from Rosenthal-von der Pütten and Krämer (2014), who categorized a wide array of robot pictures into clusters based on their perceived uncanniness (among other attributes). We selected four robot pictures with increasing levels of uncanniness, and additionally included a picture of a human male with similar image dimensions. Thus, we ended up with five pictures (agents; see Fig. 2 for pictures). Our participants were randomly assigned to evaluate one of these agents on the following characteristics, on a Likert 1 (Not at all) to 7 (Very)

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

scale: *Is this agent 1) disgusting, 2) pleasant, 3) scary, 4) trustworthy, 5) repulsive, 6) creepy, 7) friendly, and 8) likable*[2]. Due to technical problems, we omitted data for ratings of "approachable".

2.5. Materials & methods of study 3b

Study 3b was a replication attempt of Study 3a in a different population, namely, Amazon Mechanical Turk (mTurk) workers living in the USA. We sought to replicate the effects in a different population from a different cultural environment to be sure that our results from Study 3a were robust.

We distributed an online-questionnaire prepared with Qualtrics via mTurk. After omitting improper responses, we ended up with 260 participants (139 males, 109 females, 1 non-binary) who completed the survey for $0.25. The mean age of the participants was 32 years ($SD = 9.7$; range = 18−69).

Besides collecting the data online, the procedure, design and materials were conceptually identical to those in Study 3a. Our participants were randomly assigned to evaluate one of the pre-selected agents on the following characteristics, on a Likert 1 (Not at all) to 7 (Very) scale: *Is this agent 1) disgusting, 2) pleasant, 3) eery, 4) trustworthy, 5) repulsive, 6) creepy, 7) friendly, and 8) likable.* We replaced the characteristic "scary" with "eery" (we did not measure evaluations of "eery" in Study 3a, because there is no simple word for it in the Finnish language).

2.6. Materials & methods of study 4a

Study 4a was exploratory in nature, and we tested whether a "familiar" robot known anecdotally to be relatively uncanny (the robot "Sonny" from the movie *iRobot*) could be made more uncanny by morphing its face with a human face. However, we did not employ an algorithmic morphing technique, but rather downloaded a pre-existing facial image known anecdotally to be very creepy: the robot Sonny with George Clooney's face. This way we could also circumvent the possibility of our uncanny robot being confused with a real human. Conversely, we also selected a robot known anecdotally to be cute: Honda's humanoid Asimo-robot.

Invitations to participate in an online survey created with Qualtrics were posted on various student organization mailing lists in the University of Helsinki. In total, 214 respondents opened the first page of the survey and 170 (N = 170; 35 males, 129 females, 6 chose not report their sex) completed the questionnaire adequately. Of

[2] In study 3a, likability was evaluated by the question (roughly translated from Finnish) "How much do you like this agent?"

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

the participants, about 84% had some level of university education while 64% described their income level as average or less. The mean age of the participants was 27.5 years (*SD* = 8.1; range = 18−58). Participants were offered a chance to enter their email address on a separate form to participate in a movie ticket raffle (5 × 15€).

After giving informed consent, the participants were randomly assigned to evaluate one of four agents: 1) a healthy human male (from the Radboud database, as in our previous studies), 2) Honda's humanoid Asimo-robot, 3) an android character "Sonny" from the movie iRobot, and 4) "iClooney", i.e., Sonny morphed together with George Clooney's face (see Fig. 3 for pictures). As in our previous studies, each participant evaluated one agent on a Likert 1 (Not at all) to 7 (Very) scale: *Is this agent 1) disgusting, 2) pleasant, 3) scary, 4) trustworthy, 5) repulsive, 6) creepy, 7) friendly, and 8) likable.*

2.7. Materials & methods of study 4b

Study 4b was a pure replication attempt of Study 4a in an American population (Amazon mTurk workers living in the USA). Similar to what was done in Studies 3a and 3b, in Study 4b we sought to replicate the effects of Study 4a in a different population from a different cultural environment, to be sure that our results were robust.

We distributed an online-questionnaire prepared with Qualtrics via mTurk. After omitting improper responses, we ended up with 395 participants (244 males, 151 females) who completed the survey for $0.85. The mean age of the participants was 33.5 years (*SD* = 10.7; range = 19−76).

The procedure, design and materials were essentially identical to those in Study 4a. Due to technical problems, the data for the attributes "Scary" and "Friendly" are omitted.

2.8. Stimulus pictures across all studies

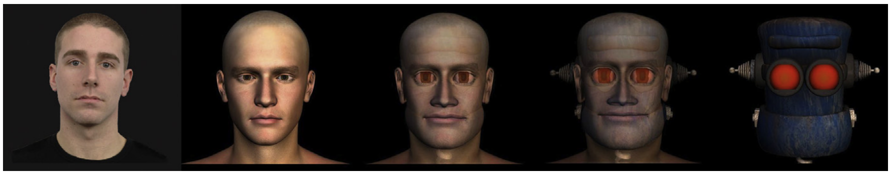

Fig. 1. Pictures of the agents used in Studies 1 and 2 (adopted from Ferrey et al., 2015). From left to right: Human (Study 2 only), Human-CGI, Human-robot, Robot-human, and Robot. The quadratic contrast "[Human + Robot] vs. [Human-robot + Robot-human]" was used to analyse differences in ratings of the following attributes: Is this agent 1) disgusting, 2) pleasant, 3) scary, 4) trustworthy, 5) repulsive, 6) creepy, 7) unapproachable, 8) friendly, and 9) nice. The human picture was obtained from the Radboud faces -database, which is a freely usable database for non-commercial scientific research (Langner et al., 2010).

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

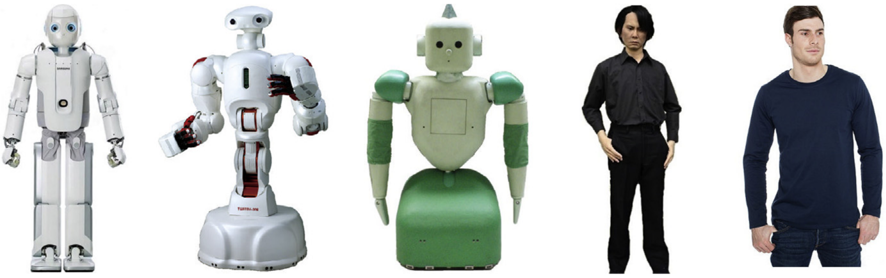

Fig. 2. Pictures of the agents used in Studies 3a and 3b. From left to right: Robot 1, Robot 2, Robot 3, Uncanny robot, Human. The following contrast coefficients were used in the analyses: 0.75 (Robot 1), 0.5 (Robot 2), 0.25 (Robot 3), -2.5 (Uncanny robot), and 1 (Human). These coefficient weights reflect the agents' *a priori* perceived uncanniness (Rosenthal-von der Pütten and Krämer, 2014), and were used to analyse differences in ratings of the following attributes: Is this agent 1) disgusting, 2) pleasant, 3) scary, 4) trustworthy, 5) repulsive, 6) creepy, 7) friendly, and 8) likable. The human picture was obtained from an online site (www.flickr.com/photos/spreadshirt/13126410163), and is under the CC BY 2.0 license.

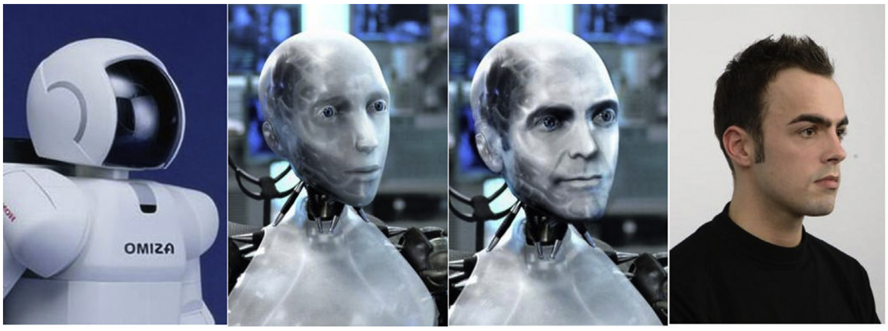

Fig. 3. Pictures of the agents used in Studies 4a and 4b. From left to right: Asimo, iRobot, iClooney, and Human. The quadratic contrast "[Human + Asimo] vs. [iRobot + iClooney]" was used to analyse differences in ratings of the following attributes: Is this agent 1) disgusting, 2) pleasant, 3) scary, 4) trustworthy, 5) repulsive, 6) creepy, 7) friendly, and 8) likable. The human picture was obtained from the Radboud faces -database, which is a freely usable database for non-commercial scientific research (Langner et al., 2010).

3. Results

3.1. Results of study 1

According to the stimulus-category hypothesis, the human-robot and robot-human morphs should both equally elicit the UV effect, whereas the non-morphed human and robot agents should not. Thus, using the general linear model ANOVA method with equal quadratic weights, we calculated the quadratic contrast "[Human + Robot] vs. [Human-robot + Robot-human]" separately for all evaluated attributes as DVs. This contrast was significant for *Disgusting* (the morphed agents were evaluated as more disgusting than the pure non-morphed human and robot agents; F(1, 217) = 4.5, B = -0.82, 95% CI [-1.58, -0.06], p = .034), *Friendly* (the human and robot agents were evaluated as more friendly $F(1, 271) = 4.9$, B = 0.81, 95% CI [0.91, 1.53], $p = .027$), and approaching significance for *Approachable* (the morphed agents were less approachable; $F(1, 217) = 3.5$, B = -0.8, 95% CI [-1.65, 0.04], $p = .061$); all other measures were non-significant (Fs < 3.5; ps = n.s.). To guard against type-1 errors we also corrected the p-values for multiple comparisons *post hoc* using the Bonferroni- and Sidak methods; after these corrections, none of the p-values were significant (ps $> .05$). See Fig. 4 for details.

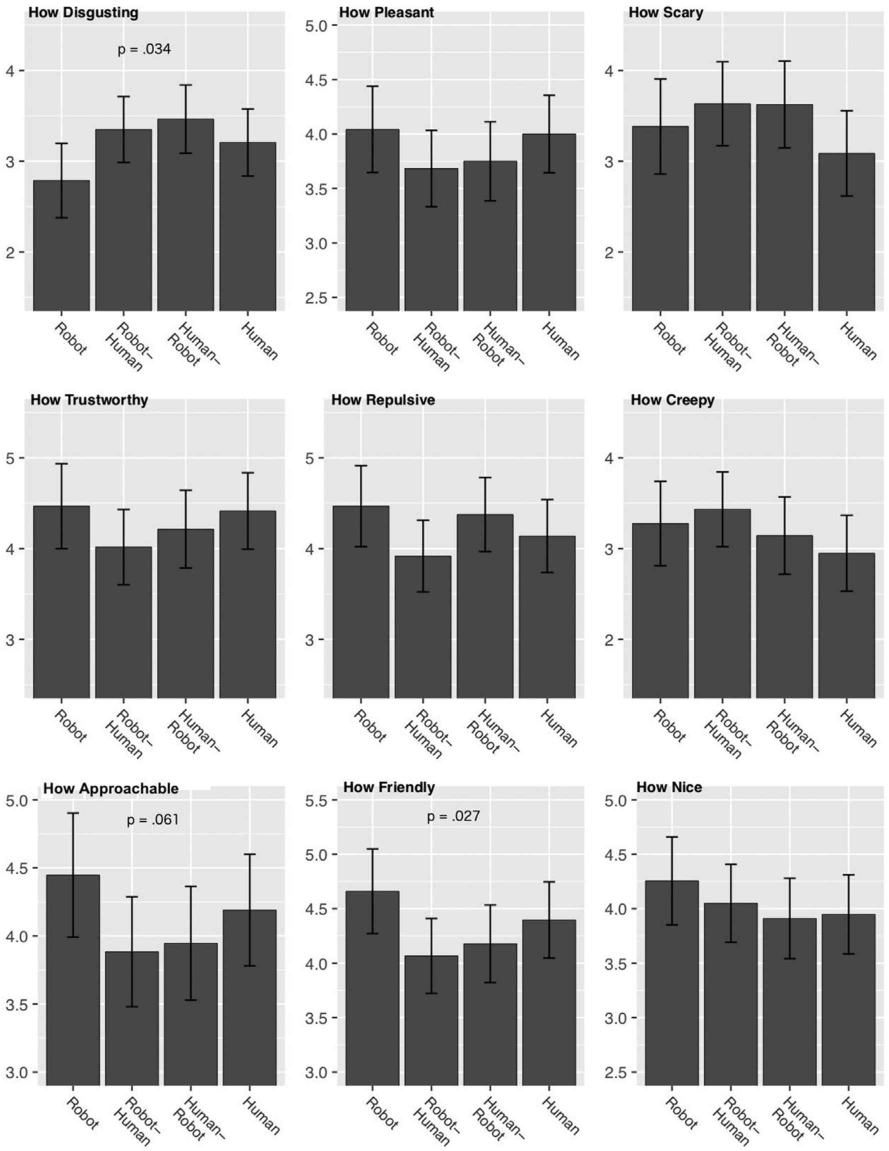

Fig. 4. Evaluations of CGI agents on different attributes in Study 1. Error bars are 95 % confidence intervals. The quadratic contrast "[Human + Robot] vs. [Human-robot + Robot-human]" is significant for *disgusting* and *friendly* and marginally significant for *approachable* (p-values shown), and non-significant for other attributes. However, after correcting for multiple comparisons, none of the effects were statistically significant.

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

The error terms were normally distributed in all models but we nonetheless reran all analyses with bootstrapped confidence intervals (using the "bias corrected accelerated" method with 1000 resamples), finding no differences in the pattern of the results. The results were also robust to adding demographic variables (age, gender, income, education) as covariates[3].

3.2. Results of study 2

We analyzed the quadratic contrast "[Human + Robot] vs. [Human-robot + Robot-human]", running analyses separately for both Human and Human-CGI agents, and all evaluated attributes as DVs. For brevity, we only report the quadratic contrasts with the Human-CGI agent (as in Study 1), and the comparison between the Human and Human-CGI agents.

The quadratic contrast was significant for ratings of *Scary* (F(1, 167) = 14.6, B = -2.3, 95% CI [-3.4, -1.1], $p < .001$)*, Repulsive* (F(1, 167) = 7.2, B = -1.56, 95% CI [-2.7, -0.41], $p = .008$)*,* and *Creepy* (F(1, 167) = 10.9, B = -2.1, 95% CI [-3.37, -0.85], $p = .001$)*,* with the morphed agents receiving higher scores than the human- or robot agents. After correcting for multiple comparisons with the Bonferroni method, the effect for *Repulsive* dropped below significance ($p = .088$). The quadratic contrast was also significant for ratings of *Disgusting* and *Pleasant,* but for these characteristics, there was no observable "valley shape" (that is, no difference between the robot and the morphed agents). The Human agent was rated as more *Human/Humane* than the Human-CGI agent (F(1, 167) = 6.06, B = 0.86, 95 % CI [0.17, 1.55], $p = .015$), but there were no other significant differences between the two human agents in the evaluated attributes (Fs < 2.5, ps = n.s.). Importantly, the significant effects were not replicated across Studies 1 and 2 for the same variables (see Fig. 5).

3.3. Results of study 3a

To test whether a "valley-shape" exists in these agents' evaluated attributes (DVs), we performed contrast analyses using the following coefficients: 0.75 (Robot 1), 0.5 (Robot 2), 0.25 (Robot 3), -2.5 (Uncanny robot), and 1 (Human). Essentially, this is a quadratic contrast comparing the human and non-uncanny robots to the uncanny robot, where the coefficients reflect the agents' *a priori* perceived uncanniness (Rosenthal-von der Pütten and Krämer, 2014).

This contrast was significant for all evaluated attributes, largely in support of the UV effect and valley shape: *Pleasant* (F(1, 120) = 21.6*,* B = 3.04, $p < .001$),

[3] We did this in all Studies with equal results, but for brevity only report it here.

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

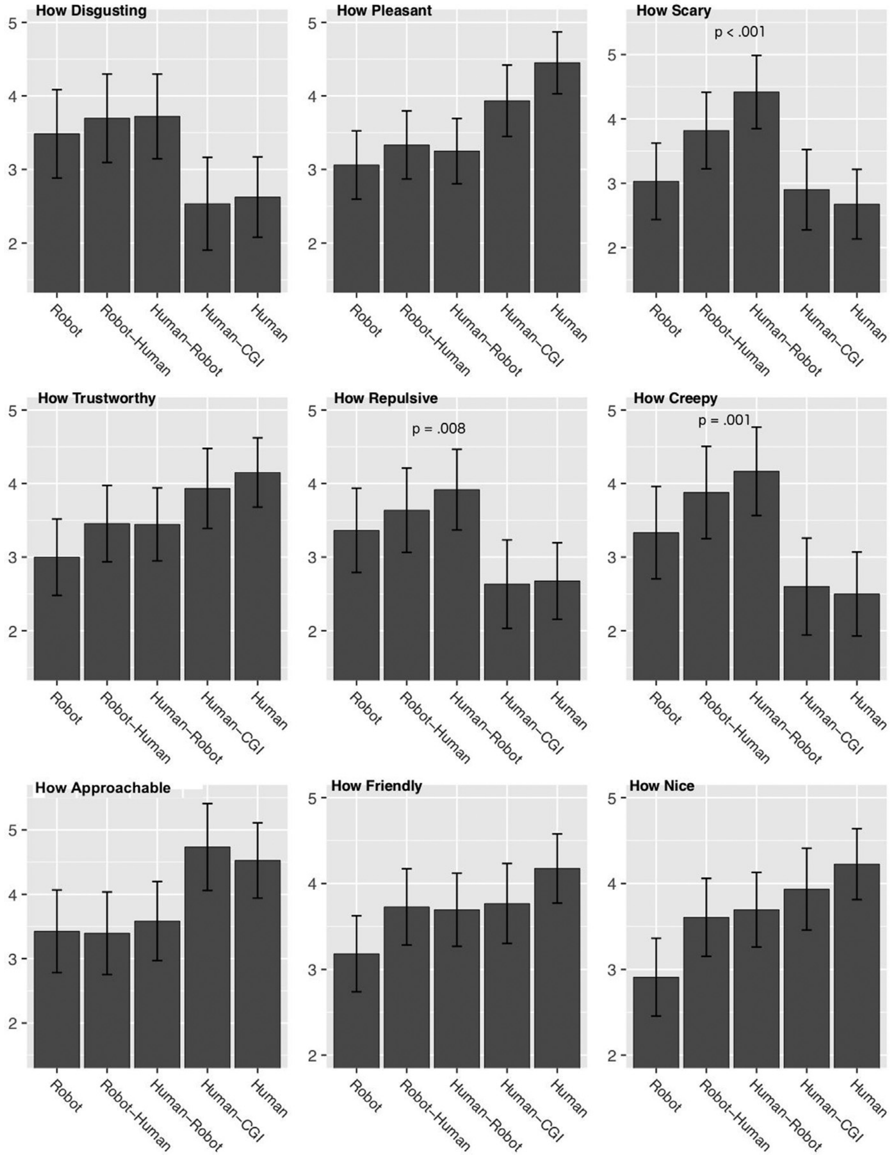

Fig. 5. Evaluations of agents on different attributes in Study 2. Error bars are 95 % confidence intervals. The quadratic contrast "[Human + Robot] vs. [Human-robot + Robot-human]" is significant for *scary*, *repulsive*, and *creepy* (p-values shown), and non-significant for other attributes. The difference between the Human and Human-CGI agents is not significant in any of the contrasted attributes.

Trustworthy (F(1, 120) = 12.9, B = 2.11, p < .001)*, Repulsive* (F(1, 120) = 23.77, B = -3.7, p < .001)*, Creepy* (F(1, 120) = 7.52, B = -2.3, p = .007)*, Friendly* (F(1, 120) = 32.6, B = 3.6, p < .001)*, Likable* (F(1, 120) = 24.5, B = 3.1, p < .001)*, Disgusting* (F(1, 120) = 6.5, B = -2.02, p = .012) and *Scary* (F(1, 120) = 8.24, B = -2.4, p = .004), all of which (with the exception of "disgusting") remained significant after correcting for multiple comparisons. See Fig. 6 for more details.

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

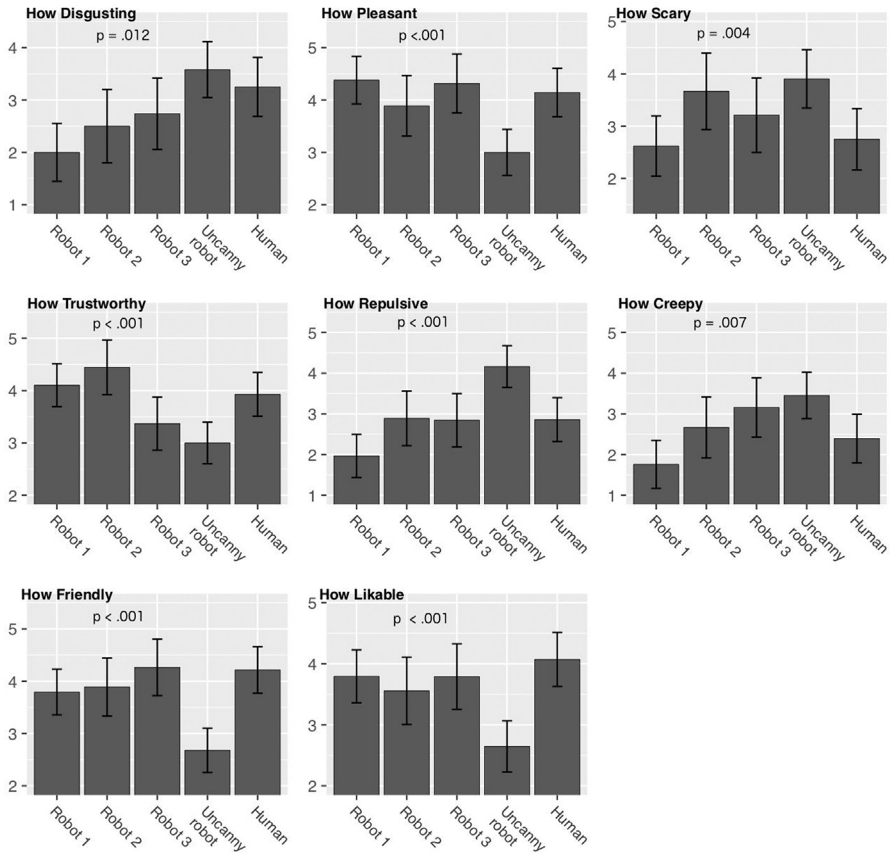

Fig. 6. Evaluations of agents on different attributes in Study 3a. Error bars are 95 % confidence intervals. The contrast using the following coefficients: 0.75 (Robot 1), 0.5 (Robot 2), 0.25 (Robot 3), -2.5 (Uncanny robot), and 1 (Human) is significant for all evaluated attributes (p-values shown).

3.4. Results of study 3b

As in Study 3a, we performed contrast analyses using the following coefficients: 0.75 (Robot 1), 0.5 (Robot 2), 0.25 (Robot 3), -2.5 (Uncanny robot), and 1 (Human). This contrast was highly significant for all evaluated attributes, largely in support of the UV effect: *Pleasant* (F(1, 255) = 148, B = 5.8, p < .001), *Trustworthy* (F(1, 255) = 63.8, B = 3.7, p < .001), *Repulsive* (F(1, 255) = 47.7, B = -4.09, p < .001), *Creepy* (F(1, 255) = 50.2, B = -5.3, p < .001), *Friendly* (F(1, 255) = 129.5, B = 5.5, p < .001), *Likable* (F(1, 255) = 109.4, B = 5.3, p < .001), *Disgusting* (F(1, 255) = 43.1, B = -3.87, p < .001) and *Eery* (F(1, 255) = 41, B = -4.5, p < .001), all of which remained significant after correcting for multiple comparisons. See Fig. 7 for more details.

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

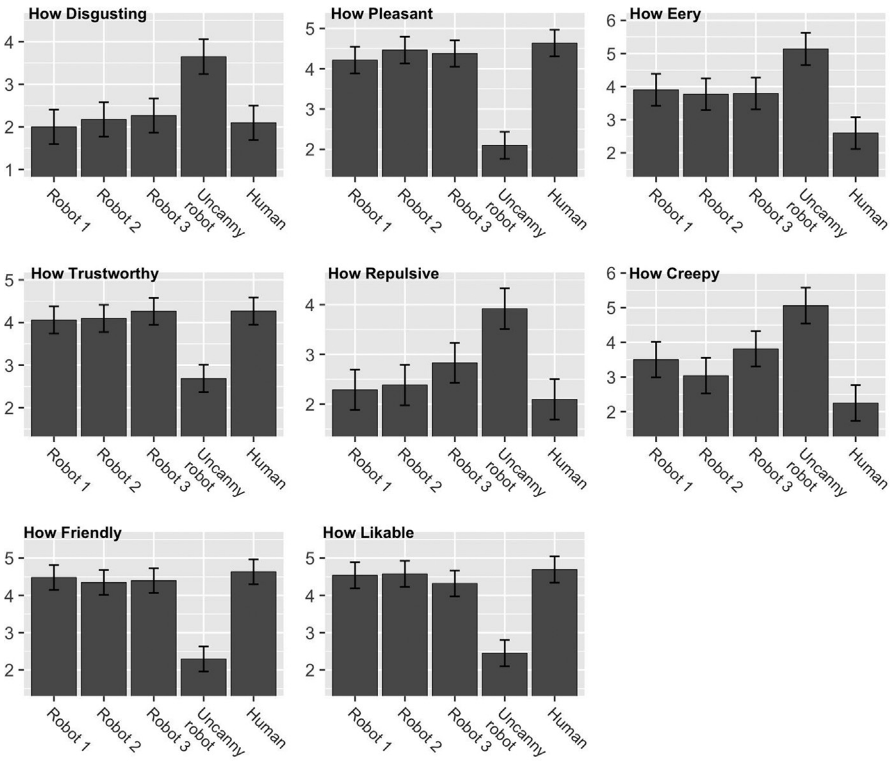

Fig. 7. Evaluations of agents on different attributes in Study 3b. Error bars are 95 % confidence intervals. The contrast using the following coefficients: 0.75 (Robot 1), 0.5 (Robot 2), 0.25 (Robot 3), -2.5 (Uncanny robot), and 1 (Human) is highly significant for all evaluated attributes (all p-values < .001).

3.5. Results of study 4a

We calculated the quadratic contrast "[Human + Asimo] vs. [iRobot + iClooney]" separately for all evaluated attributes as DVs. In support of the UV effect and valley shape, this contrast was significant for *Creepy* (F(1, 167) = 14.1, B = -1.97, $p <$ .001), *Likable* (F(1, 167) = 6.5, B = 1, p = .01), *Pleasant* (F(1, 167) = 18, B = 1.76, $p <$ .001), *Disgusting* (F(1, 167) = 17.3, B = -2.03, $p <$ .001), *Scary* (F(1, 167) = 5.35, B = -1.1, p = .02), *Trustworthy* (F(1, 167) = 6.65, B = 1.05, p = .01), *Repulsive* (F(1, 167) = 13.22, B = -1.8, $p <$ .001), but not for *Friendly* (F(1, 167) = 1.49, B = 0.48, p = .2). See Fig. 8 for more details. Most of the effects were robust for Bonferroni correction for multiple comparisons.

3.6. Results of study 4b

As in Study 4a, we calculated the quadratic contrast "[Human + Asimo] vs. [iRobot + iClooney]" separately for all evaluated attributes as DVs. In support of the UV effect and valley shape, this contrast was significant for *Creepy*

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

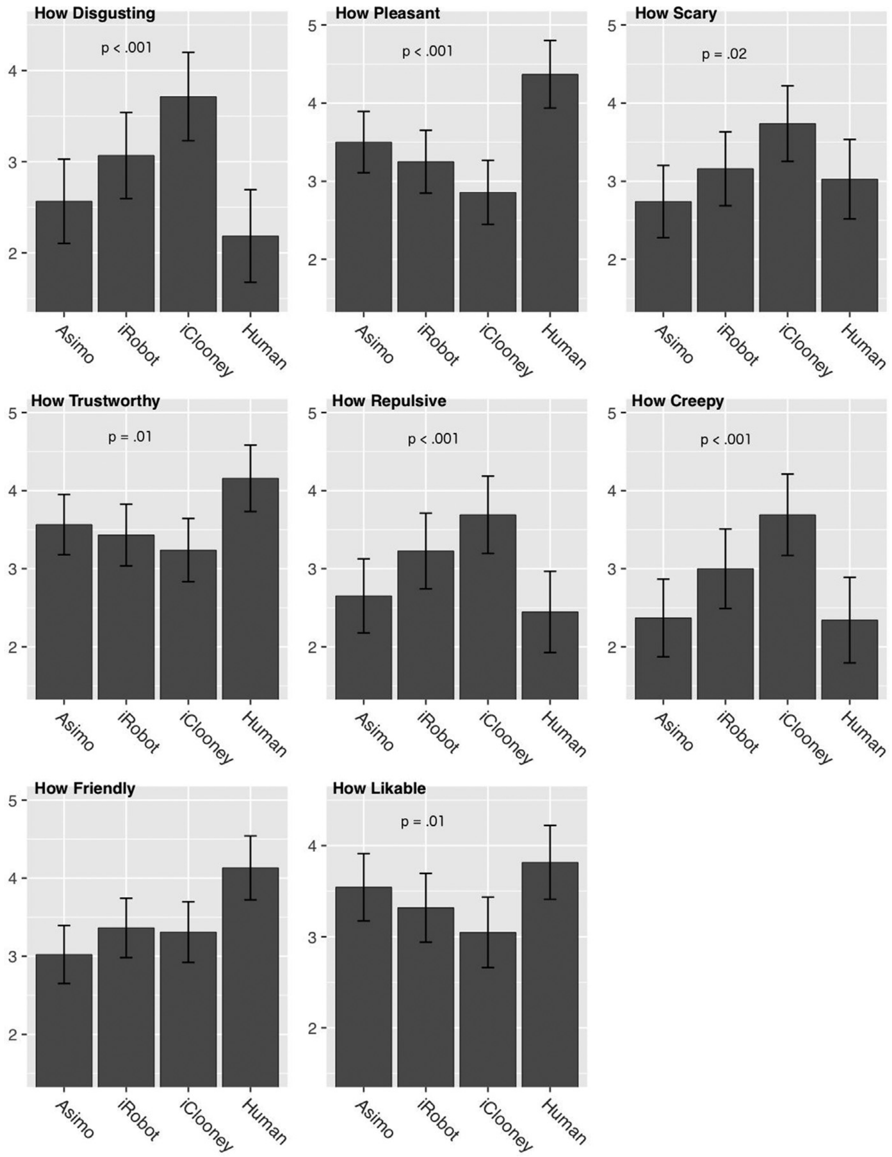

Fig. 8. Evaluations on agents on different attributes in Study 4a. Error bars are 95 % confidence intervals. The quadratic contrast "[Human + Asimo] vs. [iRobot + iClooney]" is significant for all other attributes besides friendly (p-values shown).

(F(1, 391) = 10.4, B = -1.18, p = .001), *Likable* (F(1, 391) = 6.04, B = 0.8, p = .01), *Pleasant* (F(1, 391) = 5.8, B = 0.83, p = .01), *Repulsive* (F(1, 391) = 7.56, B = -0.97, p = .006), but not for *Disgusting* (F(1, 391) = 0.006, B = 0.02, p = .9) and *Trustworthy* (F(1, 391) = 3.26, B = 0.62, p = .07). The effects for "Creepy" and "Repulsive" were robust for Bonferroni correction for multiple comparisons. See Fig. 9 for more details.

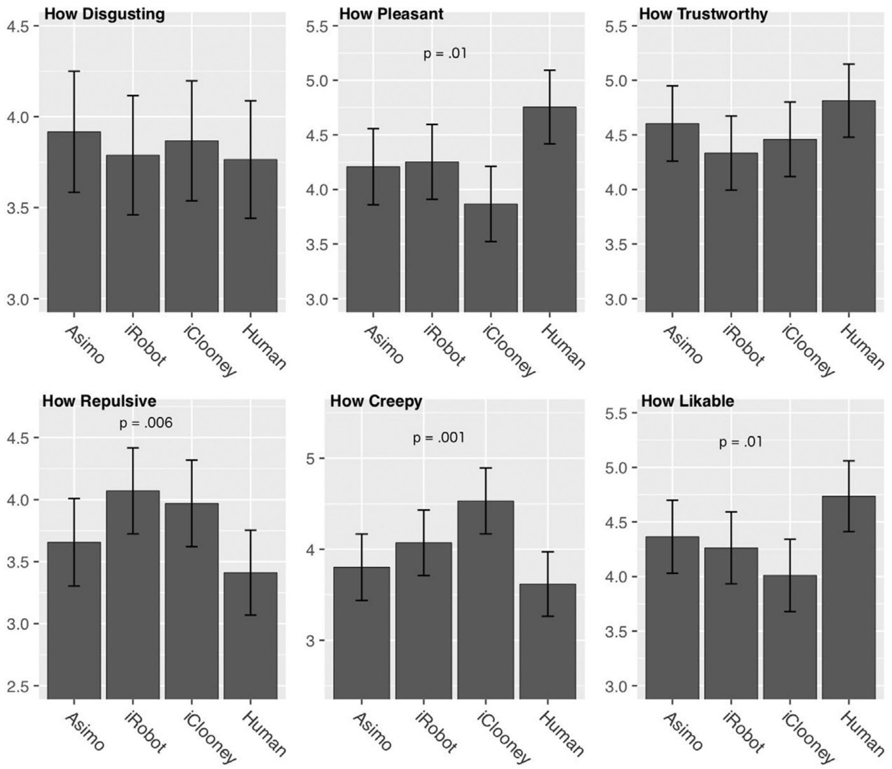

Fig. 9. Evaluations on agents on different attributes in Study 4b. Error bars are 95 % confidence intervals. The quadratic contrast "[Human + Asimo] vs. [iRobot + iClooney]" is significant for all other attributes besides *disgusting* and *trustworthy* (p-values shown).

4. Discussion

4.1. Discussion of study 1

Our results suggested that the human-robot and robot-human morphs, compared with the human and robot agents, were evaluated as more disgusting, less friendly, and marginally less approachable. However, the effects were weak, and their significance dropped below conventionally acceptable levels after correcting for multiple comparisons. Notably, we were unable to find the UV effect for the characteristics typically associated with it, namely, likability ("pleasant" and "nice"[4]) and creepiness ("creepy", "repulsive", "scary"). Moreover, the ratings for the *human* agent were relatively high for disgustingness and repulsiveness, which suggested that the computer-generated human face might not have been realistic enough, or human-like, for our participants. This implies that morphed graphics below photo-realistic levels might not be adequate to reliably induce the UV effect (which is in line with MacDorman and Chattopadhyay, 2016).

[4] Note that in Finnish, there is no word that easily translates into "likable". The closest options we chose were "miellyttävä" (pleasant) and "mukava" (nice).

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

4.2. Discussion of study 2

In Study 2, like in Study 1, the UV effect could *not* be consistently elicited. In Study 1, we observed a marginal UV effect for ratings of how disgusting, friendly, and approachable the agents were; but in Study 2 there were no effects regarding these characteristics. In Study 2, the ratings of how scary, repulsive, and creepy the agents were did elicit an effect, which was robust after controlling for multiple comparisons. However, even in these characteristics the difference between the robot and the morphed agents was not particularly large. Moreover, we did not observe significant effects for ratings of likability (how pleasant or nice the agent is); implying that these materials are not adequately robust for reliably replicating the UV effect.

4.3. Discussion of studies 3a and 3b

In Studies 1 and 2 the stimuli had been previously generated by algorithmically morphing two CGI-pictures (a human and a robot), whereas in Studies 3a and 3b the stimuli were selected, without morphing, from a large pool of photorealistic robot pictures with varying appearances. In Studies 3a and 3b, unlike Studies 1 and 2, we observed the UV effect for both positive ("likable" and "friendly", "pleasant") and negative ("creepy", "repulsive") attributes. While our results should not be viewed as direct evidence against the stimulus-category hypothesis, we nonetheless showed that the UV effect seems to be much more robust with pre-selected robot pictures than with non-photorealistic morphed ones. These results also highlight the importance of having actually photorealistic pictures to bring out the UV effect; which is in line with our results in Study 2, where the actual human agent was rated as more human/humane than the computer generated human agent. Our studies also seem to be in alignment with a recent study by MacDorman and Chattopadhyay (2016), who concluded that human-like realism is implicated in the UV effect.

Indeed, the uncanny robot in our study was very humanlike, and could in fact have been indistinguishable from a real human being (despite looking "uncanny"; see Fig. 8). Another issue relating to the stimulus materials in Studies 2 and 3a-b is that the whole body of the evaluated agents is visible. We were not certain if this was an important factor in eliciting the UV effect, and decided to focus on the agents' facial area in our following studies.

4.4. Discussion of studies 4a and 4b

In Studies 4A and 4B we tested intuitively selected images that seemed to be *prima facie* suitable for inducing the UV effect, and found the effect for most of our variables in both Studies. Importantly, the effect was prominent for the two variables that matter the most, that is, *creepy* and *likable*. We thus conclude that we were able to reliably induce the UV effect in different populations.

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

Across our Studies, it seems that the simplest way for reliably inducing the UV effect with still images is to use the series of pictures as in Studies 4a and 4b. These four images were pre-selected, had obvious *face validity,* one of which made use of the morphing technique; thus, Studies 4a and 4b combine the previous methods tested in Studies 1−3. What seems to be crucial is the need to include an actual photorealistic image of a human, and real or realistic images of robots in the stimulus stack. Moreover, our results indicate that full body images are not a requirement for inducing the UV effect.

4.5. Overall discussion

In a series of six studies we tested three sets of stimulus materials designed to induce the UV effect. In Studies 1 and 2 we tested a theory-driven approach of using images generated with morphing techniques, but failed to replicate the effects reported by Ferrey and colleagues (2015). Next, in Studies 3a and 3b, our stimuli were full body images of actual robots built in the past, which had been classified by Rosenthal-von der Pütten and Krämer (2014) by their various degrees of uncanniness. With this set of images, alongside an additional photograph of a human, we did succeed in observing a prominent UV effect. We tested these images in online and controlled laboratory environments in two different cultures, which strengthens our confidence with respect to their future use. In Studies 4a and 4b we combined both approaches and successfully found a minimal set of images that induced the UV effect also in two different cultures; and we ruled out the possibility of Studies 1 and 2 having failed due to being conducted using facial image stimuli. As far as we know, this is the first paper attempting to validate and independently replicate existing various uncanny valley -related stimulus materials in circulation in online and laboratory environments

Our results from Studies 1 and 2 are not direct evidence against the stimulus-category hypothesis. However, the results still demonstrate that it is difficult to replicate the UV effect with non-photorealistic images generated with the morphing-technique, which could be sensitive to context effects produced by within-subjects designs where participants see all images and are able to mentally intercompare them. In contrast, we did observe a prominent UV effect in Studies 3a-b and 4a-b using different types of photorealistic stimulus materials (full body and facial images). Our inability to replicate the hypothesized effects in Studies 1 and 2 is unlikely to be related to confounding factors in the laboratory environment due to our adherence to good laboratory standards, and given that Studies 3a-b and 4a-b were replicated both online and in a controlled laboratory environment.

Our results generally relate to methodological choices in experimental materials to reliably trigger the UV effect; and in particular, we showed that the UV effect could be reliably triggered with a simple set of four images that were *prima facie* salient to induce the effect. The face validity of stimulus materials likely plays an important role in designing experiments to study high-level perceptual cognitive mechanisms,

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

such as those involved in evaluating uncanny agents across numerous attributes. Our findings also suggest that to reliably elicit the UV effect in different contexts the stimulus images need to be seemingly photorealistic.

Like all studies ours also faces a number of limitations. Studies 1 and 2 were *conceptual* replications of the original study by (Ferrey et al., 2015), since we only employed a subset of the stimuli used in the original research and a between-subjects (instead of within-subjects) experimental design. However, if we failed to observe the UV effect in Studies 1 and 2 because our image palette lacked granularity, then we ought to have obtained null results also in our subsequent studies using comparable numbers of images. This was, however, not the case: across our subsequent studies we did observe prominent UV effects and showed that computer generated images probably need to be highly realistic to reliably induce it (see also the previous research by MacDorman and Chattopadhyay, 2016).

Statistical between-subjects models, compared with within-subjects models, typically have larger error terms since subject variability across conditions cannot be modeled (in statistical terms, subject is nested within condition). However, between-subjects designs allow for collecting larger sample sizes with high statistical power and fewer resources given the much shorter time required for participants to complete the task, compared with counter-balanced within-subjects designs. Long experiments also entail serious data quality issues such as participant fatigue, boredom, losing naiveté, and demand characteristics such as being able to guess the study hypothesis. Thus, shorter lasting between-subjects experiments yield higher quality (less noisy) data than many longer lasting within-subjects experiments. Should the materials used by Ferrey and colleagues to test the stimulus category hypothesis be robust against contextual variation, they should produce similar results also in a between-subjects experiment. Indeed, our studies suggest that between-subjects designs should preferably be used to find reliable stimulus materials for studies on the UV effect, and to robustly test for possible novel hypotheses.

Our study populations were not fully representative of the general population; however, this limitation is mitigated by 1) having collected data from two countries with different cultures, and by 2) collecting data both online and offline — yet still finding converging results repeatedly. Moreover, collecting data in a public library (as well as mTurk) enabled us to obtain a more heterogeneous sample compared with typical experiments in psychological science, where samples typically consist of 20—25-year-old female students. Finally, in all our studies our participants responded to 6—9 dependent variables (attributes for the evaluated images) using a seven point Likert scale. The morphed images used in Studies 1 and 2 might be sensitive to a 0.01 -pointed visual-analog rating scale (ranging from 0 to 1; i.e. a slider-scale), as used by Ferrey and colleagues (2015). However, since we did observe positive results using Likert scales in Studies 3a and 3b, it is unlikely that observing the

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

UV effect depends on participants using a visual-analog slider scale to rate their responses; and even if observing the UV effect were limited to this methodological issue, other researchers should be aware of it.

Finally, the uncanny robot image used in Study 3 had a seemingly angry facial expression, which might have confounded participants' ratings on its characteristics; and in Study 4, it is unclear why the morphed image "iClooney" elicited strong UV effects while the other morphed images (in Studies 1 and 2) did not (see also Mäkäräinen et al., 2015). We propose that this might be related to relying solely on CGI, and thus missing photorealism, when creating morphed images. Furthermore, in Study 2 we found that the CGI human image was indeed rated as more disgusting than an actual human face; and in Study 4 we contrasted photorealistic images with CGI images to shed light on the observed effects. Obviously, the CGI images are a "lower resolution" representation of a given entity compared with actual photographs.

Until now most UV research has primarily been focused on understanding the phenomenon on a basic level, and trying to explain what gives rise to it. For future studies we suggest studying the UV effect in applied settings so that its impact on human cognition can be assessed beyond merely perceptual mechanisms. For example, it would be interesting to evaluate how uncanny agents are perceived morally, or how they are treated in game theoretical games as opponents. Along these lines, future research could also evaluate how the uncanniness of an agent influences their perceived trustworthiness in business transactions or in customer service.

In conclusion, we studied three different sets of stimulus materials created for studying the UV phenomenon and ruled out several confounding factors as possible explanations for our results. We found the UV effect using a previously published set of five images, and with a simple set of four images selected by ourselves. However, we could not replicate the effects for the stimuli generated with a morphing technique (Ferrey et al., 2015). The stimulus-category hypothesis might still be correct, although our results cast some doubt on it. Future studies need to try to find a robust way of inducing the UV effect in different conditions and contexts with simple and economic survey tools, and in a way that can be independently replicated by other research teams. We also suggest that future studies making use of the UV effect should further try to replicate and validate stimulus materials across different social or behavioral contexts, in both basic and applied settings.

Declarations

Author contribution statement

Jussi Palomäki, Michael Laakasuo: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

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

Anton Kunnari, Marianna Drosinou: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Mika Koverola, Noora Lehtonen, Juho Halonen, Marko Repo: Performed the experiments; Wrote the paper.

Funding statement

This work was supported by Jane and Aatos Erkko Foundation and Weisell-foundation (Project: Moralities of Intelligent Machines).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Data associated with this study has been deposited on figshare, at URL https://doi.org/10.6084/m9.figshare.7304189.v1.

Acknowledgements

This project would not have been possible without a generous grant provided by Jane and Aatos Erkko Foundation to study the moral interactions between humans and AI (www.moim.fi and www.jaes.fi). MD would like to give special thanks to Tiina and Antti Herlin foundation for supporting her PhD work, during which she could collaborate with the 'Moralities of Intelligent Machines' research team. ML would also like to thank Apophenia for providing inspiration. All authors are members of the Moralities of Intelligent Machines research team (www.moim.fi).

References

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.

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.

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.

Green, R.D., MacDorman, K.F., Ho, C.C., Vasudevan, S., 2008. Sensitivity to the proportions of faces that vary in human likeness. Comput. Hum. Behav. 24 (5), 2456–2474.

Hanson, D., 2006. Exploring the aesthetic range for humanoid robots. In: Proceedings of the ICCS/CogSci-2006 Long Symposium: toward Social Mechanisms of Android Science, Vancouver, Canada, pp. 16–20.

International Federation of Robotics, 2016. World Robotics Report 2016.

Kätsyri, J., Förger, K., Mäkäräinen, M., Takala, T., 2015. A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Front. Psychol. 6, 390.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.D., 2010. Presentation and validation of the radboud faces database. Cognit. Emot. 24 (8), 1377–1388.

MacDorman, K.F., 2005, July. Androids as an experimental apparatus: why is there an uncanny valley and can we exploit it. In: CogSci-2005 workshop: toward social mechanisms of android science, Vol. 106118.

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., Green, R.D., Ho, C.-C., Koch, C.T., 2009. Too real for comfort? Uncanny responses to computer generated faces. Comput. Hum. Behav. 25, 695–710.

Mäkäräinen, M., Kätsyri, J., Förger, K., Takala, T., 2015. The funcanny valley: a study of positive emotional reactions to strangeness. In: Proceedings of the 19th International Academic Mindtrek Conference. ACM, pp. 175–181.

Matsuda, Y.T., 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., rsbl20120346.

Mitchell, W.J., Szerszen, K. A. Sr., Lu, A.S., Schermerhorn, P.W., Scheutz, M., MacDorman, K.F., 2011. A mismatch in the human realism of face and voice produces an uncanny valley. Iperception 2, 10–12.

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

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251) aac4716-aac4716.

Rosenthal-von der Pütten, A.M., Krämer, N.C., 2014. How design characteristics of robots determine evaluation and uncanny valley related responses. Comput. Hum. Behav. 36, 422–439.

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

Yamada, Y., Kawabe, T., Ihaya, K., 2012. Can you eat it? A link between categorization difficulty and food likability. Adv. Cognit. Psychol. 8, 248–254.

Yamada, Y., Kawabe, T., Ihaya, K., 2013. Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. Jpn. Psychol. Res. 55 (1), 20–32.

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