Computers in Human Behavior 103 (2020) 21–30

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley

Mathur Maya B.[a,b,*], David B. Reichling[c], Francesca Lunardini[d,1], Alice Geminiani[d,1], Alberto Antonietti[d,1], Peter A.M. Ruijten[e,1], Carmel A. Levitan[f,1], Gideon Nave[g,1], Dylan Manfredi[g,1], Brandy Bessette-Symons[h,1], Attila Szuts[i,1], Balazs Aczel[i,1]

[a] Department of Epidemiology, Harvard University, Boston, MA, USA
[b] Quantitative Sciences Unit, Stanford University, Palo Alto, CA, USA
[c] University of California at San Francisco (ret.), San Francisco, CA, USA
[d] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy
[e] Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, Netherlands
[f] Department of Cognitive Science, Occidental College, Los Angeles, CA, USA
[g] Marketing Department, The Wharton School of the University of Pennsylvania, Philadelphia, PA, USA
[h] Department of Psychology, Ithaca College, Ithaca, NY, USA
[i] Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

ARTICLE INFO

Keywords:
Uncanny valley
Human-robot interaction
Psychology
Social interaction
Categorical perception

ABSTRACT

Android robots that resemble humans closely, but not perfectly, can provoke negative feelings of dislike and eeriness in humans (the "Uncanny Valley" effect). We investigated whether category confusion between the perceptual categories of "robot" and "human" contributes to Uncanny Valley aversion. Using a novel, validated corpus of 182 images of real robot and human faces, we precisely estimated the shape of the Uncanny Valley and the location of the perceived robot/human boundary. To implicitly measure confusion, we tracked 358 participants' mouse trajectories as they categorized the faces. We observed a clear Uncanny Valley, though with some interesting differences from standard theoretical predictions; the initial apex of likability for highly mechanical robots indicated that these robots were still moderately dislikable, and the Uncanny Valley itself was positioned closer to the mechanical than to the human-like end of the spectrum. We also observed a pattern of categorization suggesting that humans do perceive a categorical robot/human boundary. Yet in contrast to predictions of the category confusion mechanism hypothesis, the locations of the Uncanny Valley and of the category boundary did not coincide, and mediation analyses further failed to support a mechanistic role of category confusion. These results suggest category confusion does not explain the Uncanny Valley effect.

1. Introduction

Android robots have rapidly entered our social sphere. We now entrust them with providing therapy to children with autism and older adults, coaching patients on health behavior change, and collaborating with astronauts in space stations (Rabbitt, Kazdin, & Scassellati, 2015; Weisberger, 2018). Yet human-robot interactions can be fraught with social peril. Robots that closely resemble humans but are not perfectly human-like can elicit unexpectedly negative emotional reactions in human viewers, jeopardizing the robots' social success. This "Uncanny Valley" theory (Mori, 1970) has dominated discussion of human reactions to anthropomorphic robots in both popular culture and research literature. Specifically, the theory posits that as android robots increasingly resemble humans, their likability improves until a point at which it abruptly drops to a negative value because the robots become dislikable and eerie (Fig. 1). As the robots' human-likeness continues to increase past this "Uncanny Valley", they again become likable and eventually become maximally likable as they become indistinguishable from humans. The Uncanny Valley does seem to occur in real android robots that were intentionally designed to interact with humans (Mathur and Reichling (2016); Slijkhuis (2017); Lischetzke, Izydorczyk, Hüller, and Appel (2017); Jung and Cho (2018)).

With android robots increasingly becoming everyday technology, there is a pressing need to understand the psychological mechanisms underlying the Uncanny Valley effect. Some hypotheses posit relatively

* Corresponding author. Quantitative Sciences Unit, Stanford University, Palo Alto, CA, USA.
E-mail address: mmathur@stanford.edu (M.B. Mathur).
[1] Order of contributing authors was randomized by site and then by author name.

https://doi.org/10.1016/j.chb.2019.08.029
Received 27 June 2019; Received in revised form 22 August 2019; Accepted 30 August 2019
Available online 18 September 2019

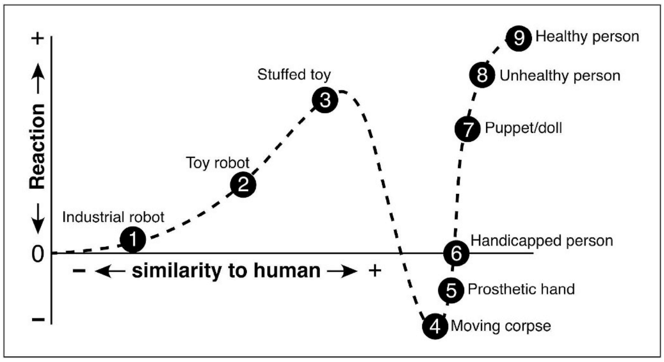

Fig. 1. The hypothesized Uncanny Valley, adapted from Mori (1970).

high-level social and affective mechanisms: robots in the Uncanny Valley might prime awareness of one's own mortality (Ho, MacDorman, & Pramono, 2008), prompt dehumanization responses similar to that directed at human targets of prejudice (Wang, Lilienfeld, & Rochat, 2015), or trigger impulses to avoid pathogens (MacDorman, Green, Ho, & Koch, 2009). In contrast, according to a longstanding lower-level hypothesis known as "category confusion", there is a deep-seated perceptual distinction between the categories "human" and "non-human", and androids that are difficult to categorize as one or the other cause aversion due to the categorization difficulty itself (Jentsch (1906, pp. 195–198); English translation: Jentsch (1997)). That is, negative reactions to robots in the Uncanny Valley may be a special case of general aversion to cognitive inhibition caused by competing perceptual representations (Ferrey, Burleigh, & Fenske, 2015; Freeman & Johnson, 2016).

Similar mechanisms may underlie confusion occurring between categories of gender, race, age, sexual orientation, attitude, attractiveness, skin tone, accent, and relationship type (Fiske & Taylor, 2008). If there is indeed a general aversion to category confusion, the category boundaries most relevant to the Uncanny Valley – those demarcating "human" vs. "non-human" and "animate" vs. "inanimate" – might also elicit an enhanced response given humans' specialized adaptions for face perception (Kanwisher, 2000). Indeed, stimuli depicting inanimate objects across a spectrum of realism, such as morphs between a photograph of a car and a computer-generated rendering of the car, seem to elicit weaker Uncanny Valley effects than stimuli depicting humans and humanoids (e.g., MacDorman and Chattopadhyay (2016)'s Fig. 5; Schwind, Leicht, Jäger, Wolf, and Henze (2018)), though stimuli depicting animals do seem to elicit Uncanny Valley effects (e.g., MacDorman & Chattopadhyay, 2016; Schwind et al., 2018).

Supporting some predictions of the category confusion explanation for the Uncanny Valley, human viewers may indeed perceive a categorical boundary between human and various near-human stimuli (Looser & Wheatley, 2010; Weis & Wiese, 2017), and mid-range stimuli seem to elicit the most confusion based on implicit measures such as reaction time and eye-gaze direction (Cheetham, Wu, Pauli, & Jancke, 2015, 2013; MacDorman & Chattopadhyay, 2016; Mathur & Reichling, 2016; Yamada, Kawabe, & Ihaya, 2013). However, it remains unclear whether confusion at the category boundary actually *causes* negative emotional responses to ambiguous faces or is merely an epiphenomenon. The mechanistic account would predict that the point on the nonhuman-human spectrum eliciting maximum confusion should coincide with the point eliciting the most negative social and affective responses; some studies have preliminarily supported this prediction (Mathur & Reichling, 2016; Yamada et al., 2013), but others have not (Cheetham et al., 2015, 2014; Looser & Wheatley, 2010; MacDorman & Chattopadhyay, 2016).

The apparent conflict between these findings may partly reflect methodological limitations (Kätsyri, Förger, Mäkäräinen, & Takala, 2015; Lay, Brace, Pike, & Pollick, 2016). First, many studies to date have used as few as three stimuli spanning only a limited, very human-like region of the nonhuman-human spectrum (approximately points 7–9 in Fig. 1). If all stimuli are more human-like than the point eliciting the most negative Uncanny Valley reactions, then affective patterns to these stimuli might increase monotonically with human-likeness rather than showing a characteristic Uncanny Valley curve (as seen in Cheetham, Suter, and Jancke (2014, 2015); Looser and Wheatley (2010); MacDorman and Chattopadhyay (2016)). Stimuli whose truncated range precludes observing the Uncanny Valley itself may also disrupt assessment of category confusion in the Uncanny Valley. Second, most studies have generated stimuli through digital image morphing, a method that risks producing unrealistic transitional images exhibiting, for example, partially transparent facial features or incompatible combinations of features that would be avoided in real-world robot design (Kätsyri et al., 2015; Kawabe, Sasaki, Ihaya, & Yamada, 2017). It is plausible that these unrealistic features themselves, rather than the stimuli's degree of human-likeness, might produce artifactual confusion or aversion. Statistical methods used in previous research have rarely assessed a mechanistic role of category confusion, instead focusing on certain downstream predictions of the category confusion mechanism hypothesis, for example by assessing the association between confusion and emotional responses. Two studies more directly assessed whether reaction time, a coarse measure of confusion, statistically mediated the relationship between human-likeness and ratings of eeriness or weirdness, but with differing findings (Carr, Hofree, Sheldon, Saygin, & Winkielman, 2017; Mathur & Reichling, 2016). Carr et al. (2017) reported mediation, but that study used only three stimuli; Mathur and Reichling (2016) did not detect mediation, but this was a secondary analysis of reaction times in a mechano-humanness rating task rather than a binary categorization task.

The present study aims to: (1) provide the most precise estimate to date of the shape of the Uncanny Valley curve in real-life robots and humans; (2) estimate the degree of human-likeness marking the perceptual category boundary between "non-human" and "human"; and (3) rigorously assess whether category confusion is a mechanism for Uncanny Valley effects. We first assembled a large corpus of face images of socially interactive robots that have actually been built. We minimized variation on potential confounders such as the face's perceived emotion (Lay et al., 2016) and ensured that the faces were well-distributed across the full spectrum of human-likeness, enabling precise estimates of the Uncanny Valley curve and of the location of the perceived boundary between the categories "robot" and "human". For improved generalizability, we recruited participants at six collaborating sites in four countries. Participants rated each face on human-likeness and likability, and they attempted to rapidly categorize each face as "robot" or "human" while we collected measures of category confusion. We assessed confusion using validated measures based on mouse-tracking (Freeman & Johnson, 2016; Mathur & Reichling, 2016) to supplement coarser existing measures based on reaction time. We assessed for statistical mediation as predicted by the mechanistic account of category confusion in a manner that accommodated the expected nonlinear relationships between human-likeness, confusion, and likability. Subsequent sections of this paper are structured as follows: we will describe methods for stimulus validation, for measurement of category confusion, and for participant recruitment, then describe statistical methods and results for each of the three aims in turn, then conclude with a general discussion.

2. Data collection methods

All methods and statistical analyses were preregistered in detail; the Supplement (Section 1.2) describes and justifies some deviations from this protocol. All measures and experiments are reported, and we determined sample sizes in advance. All data, materials, analysis code, and the preregistration are publicly available and documented (https://osf.io/2v6f4/).

2.1. Face stimuli

We selected stimuli depicting the faces of real robots designed for social interaction, as well as faces of real humans. To support assessing a potential role of category confusion in mediating the relationship between human-likeness and likability, we attempted to select stimuli in a manner that would minimize confounding of the relationships between human-likeness, confusion, and likability (VanderWeele, 2015). For example, if displaying more positive emotion causes a face to be perceived as more human-like and also causes the face to be perceived as more likable, then perceived emotion could act as a confounder that would compromise causal conclusions from mediation analysis (VanderWeele, 2015).

We identified face images using an objective Internet search process similar to that of Mathur and Reichling (2016). In addition to applying Mathur and Reichling (2016)'s inclusion and exclusion criteria (reproduced in the Supplement, Section 1.1), we applied the following inclusion criteria to further minimize variation on potential confounders. First, the faces had to be photographed in frontal view. Second, the faces had to be perceived as displaying neutral emotion, defined as having a mean rating between −20 and + 20 on a visual analog scale ranging from −100 to +100. Third, the faces had to be densely spread over the entire spectrum from extremely mechanical to extremely human-like, rather than concentrated in only certain parts of the spectrum. Specifically, on a continuous scale of "mechano-humanness" (MH) score ranging from −100 ("extremely mechanical") to +100 ("extremely human-like"), we required that there be no 50-point span of MH score occupied by fewer than 20 faces.

To collect face stimuli, we first reviewed the existing set of 80 robot faces from Mathur and Reichling (2016); these faces had existing estimates of MH score. We discarded images that failed the more stringent inclusion criteria used here (e.g., because the photo was a 3/4 view or the face displayed too much emotion). For failed images, we searched the Internet to try to identify alternative photos of the excluded robot that did meet the criteria. The stimuli in Mathur and Reichling (2016) were somewhat limited by their sparse coverage of the very human-like range (i.e., MH score above about +90); to resolve this limitation in the present stimuli, we also obtained images of actual human faces from www.shutterstock.com, an online image bank. To obtain these human faces, we performed a broad search using terms such as "portrait", "face", and "unemotional", attempting to select an assortment of faces meeting the same visual criteria as used for the robot faces and spanning a range from unmistakably human to somewhat artificial in appearance. The faces appearing unmistakably human included those that seemed unusually "imperfect" due, for example, to particularly prominent or asymmetric facial features. The faces appearing somewhat artificial included those that seemed unusually "perfect" due, for example, to symmetry, lack of flaws, or heavy use of makeup.

To ensure that the resulting robot and human faces were adequately densely distributed throughout the MH spectrum, we iterated between finding additional candidate images that preliminarily appeared to meet the inclusion criteria and testing the candidate images for perceived emotion and MH score using groups of pilot participants. At least 26 pilot participants (mean number of participants: 40) rated each candidate face on perceived emotion and MH score. We ultimately included 182 faces, comprising 122 robots and 60 humans (Fig. 2). The final validated corpus of images, along with summary measures of their ratings on all analyzed variables in this study, is publicly available for by-attribution use in future research (https://osf.io/2v6f4/). The faces had mean MH score −12.3 (−53.5 for the robots and 71.5 for the humans) and had mean likability −5.4 (−32.9 for the robots and 50.5 for the humans).

2.2. Measures of human-likeness, confusion, and likability

We measured human-likeness by asking, "How mechanical versus human-like does this face look?"; participants responded using a bipolar visual analog scale ranging from −100 to +100 with the endpoints labeled "extremely mechanical" and "extremely human-like" (Mathur & Reichling, 2016). We refer to this measure as "mechano-humanness" or "MH" score. We measured likability by asking participants to "Estimate how friendly and enjoyable (or creepy) it might be to interact with the robot in some everyday situation, such as asking a question at a museum's information booth"; participants responded on a similar visual analog scale with the endpoints labeled "Less friendly; more unpleasant and creepy" and "More friendly and pleasant; less creepy" (Mathur & Reichling, 2016).

We used validated open-source software (Mathur & Reichling, 2019) to collect five established measures of category confusion (e.g., Freeman, Ambady, Rule, and Johnson (2008)). Participants viewed the faces sequentially and were asked to rapidly categorize each face as "robot" or "human" by clicking on one of two buttons presented on the left and right sides of the window (Fig. 3). (Methodological details of the categorization task are available in Mathur and Reichling (2019).) Ambiguous stimuli are thought to activate mental representation of both categories simultaneously, leading to dynamic competition that manifests in real time as unstable mouse trajectories (Freeman & Johnson, 2016). That is, because the participant is continuously or alternately attracted to both categories, the mouse trajectory may contain frequent direction changes and may diverge substantially from a direct path from the start position to the location of the category button ultimately chosen.

Therefore, similarly to Freeman et al. (2008), as primary measures of confusion, we collected as primary measures of confusion: (1) the number of times the participant's mouse changed directions horizontally during categorization (*x-flips*); (2) the maximum horizontal deviation between the participant's mouse trajectory and an ideal trajectory consisting of a straight line from the participant's initial cursor position to the finally chosen category button (*maximum x-deviation*; red solid line in Fig. 3); and (3) the *area* between the ideal and actual trajectories (pink shading in Fig. 3). We additionally measured: (4) the *peak speed* of the participant's cursor (ambiguous stimuli tend to produce higher peak speeds, reflecting abrupt category shifts (Freeman, Pauker, & Sanchez, 2016)); and (5) the total *reaction time* for the trial (ambiguous stimuli tend to produce longer reaction times). Because the latter two measures have limitations as measures of category confusion (Freeman et al., 2016), we made an *a priori* decision to designate them as secondary measures. Each participant provided measures of MH score, confusion, and likability in that order for all 182 faces. To minimize the possibility of task interference or memory effects due to a participant's repeated exposure to each face, we collected each measure in a separate wave of data collection spaced by approximately one week. Within each wave, the order of the faces was randomized for each participant. At the end of the first wave, participants also completed basic demographic measures of age, sex, education level, and race/ethnicity. At the end of each wave, participants reported any technical or comprehension problems.

2.3. Participants

We collected data at six colleges and universities in the United States, Hungary, the Netherlands, and Italy; we recruited the data collection sites through the authors' previous collaborations and the online platform StudySwap (https://osf.io/meetings/StudySwap). Supplement Table S1 describes characteristics of the sites. Aggregating across sites, we analyzed data from 358 participants, who were 72% female with mean age 21.5 years; further demographic characteristics are described in Supplement Table S2. Each site aimed to collect data on at least 50 English-speaking participants in a quiet lab or classroom on lab-provided computers that were pre-tested for accurate collection of mouse-tracking data. Labs incentivized participation using various monetary compensation, course credit, or volunteer schemes. All participants completed the study using the same Qualtrics questionnaires provided by the lead authors, MBM and DBR. Each lab secured its own ethics approval or waiver as appropriate to its location. We determined sample sizes in advance based on considerations detailed in the preregistered protocol (https://osf.io/mu5xj/registrations).

Based on *a priori* criteria, we excluded participants who did not complete all three waves of the questionnaire, whose data indicated technical problems (e.g., reflecting rare, idiosyncratic timing issues that caused no times to be recorded for a participant, or caused timing to stop prematurely;

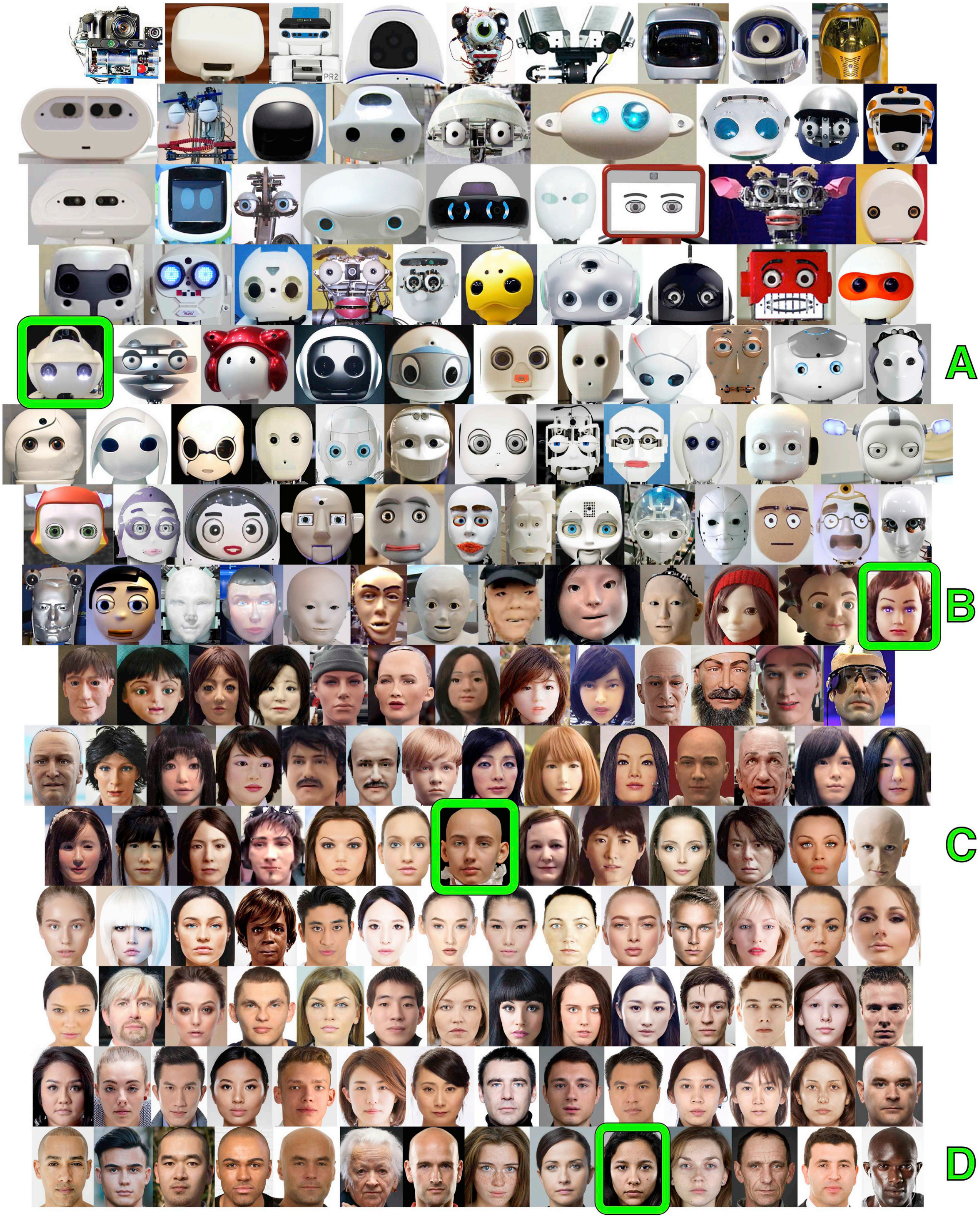

Fig. 2. Robot and human face stimuli displayed in ascending order of mean mechano-humanness (MH) score. Boxed faces are those with MH scores closest to the MH scores associated with: (A) the initial likability apex of Uncanny Valley curve (estimation described in Section 3); (B) the likability low point of Uncanny Valley; (C) the robot/human category boundary (estimation described in Section 4); and (D) the final apex of likability.

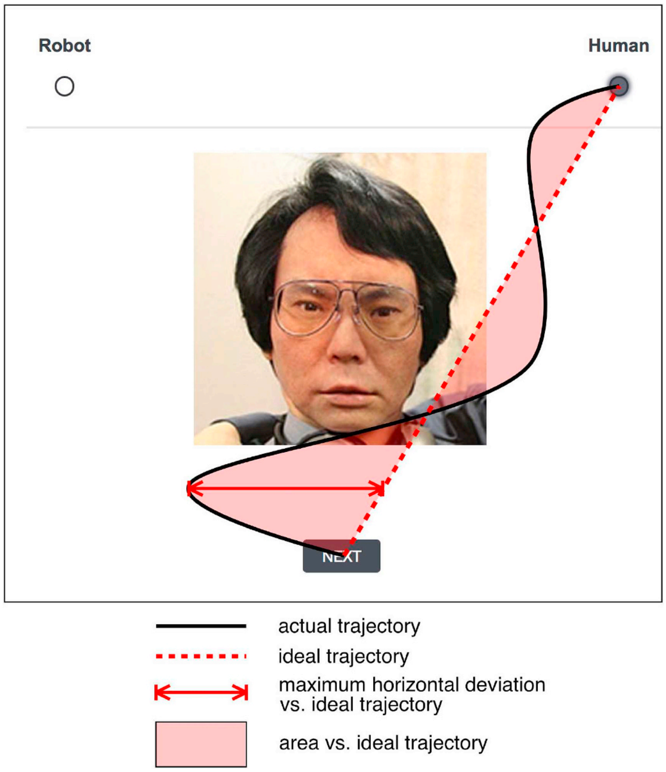

Fig. 3. Mouse-tracking measures of category confusion (reproduced from Mathur and Reichling (2019)).

Mathur and Reichling (2019)), or for whom there were known data collection errors (e.g., the waves of data collection were run in the wrong order). Supplement Fig. S1 details the number of participants excluded for each reason. Additionally, we excluded individual trials in which the value of any confusion measure was larger than its 75th percentile plus 1.5 times its interquartile range or smaller its 25th percentile minus 1.5 times the interquartile range; we made this decision because the confusion measures could potentially take on extreme values if, for example, a participant made an uncontrolled cursor movement. We did not exclude trials in which the participant chose the wrong category for the face (e.g., selected "human" when presented with a robot) because we expected many faces to be quite hard to judge. After all exclusions, the analysis dataset comprised 358 participants, totalling 55,430 ratings of the 182 faces.

3. Estimating the shape of the Uncanny Valley

3.1. Statistical methods

Throughout, one author (MBM) performed all statistical analyses in R (version 3.5.1). All analyses described in the main text were conducted with means by face as the unit of analysis[2]. In all analyses, we adjusted for a face's mean perceived emotion rating, as estimated during stimulus validation, because we suspected that emotion might statistically confound the relationship between MH score and likability. We first estimated the Uncanny Valley curve by fitting ordinary least squares models that regressed likability on polynomial terms for MH score (e.g., MH, MH^2, MH^3, etc.). We mean-centered MH score in analysis, but we report and plot results on the uncentered scale for interpretability, except where otherwise noted. We used Akaike's Information Criterion (AIC) to select the lowest-order and best-fitting polynomial model (Akaike, 1974). We weighted each data point by its inverse-variance of likability to account for the fact that some faces were rated with more precision than others, although unweighted analyses yielded very similar results (Supplement, Section 2.3).

3.2. Results

Fig. 4 shows the best-fitting and most parsimonious model for the relationship between MH score and likability, which was a six-degree polynomial in MH score. As predicted by the Uncanny Valley theory, estimates from this model indicated that as faces progressed from extremely mechanical (MH score near −100) to somewhat less mechanical, likability tended to increase to a point, reaching an initial apex of −18.0 for faces with an MH score of −80.9. After this initial apex, as faces continued to become more human-like, likability began to decrease, dropping to its overall lowest point of −67.4 for faces with an MH score of −23.6. Beyond this Uncanny Valley, as faces continued to become more human-like, their likability once again tended to increase monotonically, ultimately reaching a maximum of 59.6 for faces nearly indistinguishable from humans (i.e., those with an MH score of 93.2). Thus, all key features of the theorized Uncanny Valley were apparent in these stimuli. As a post hoc analysis suggested during peer review, we conducted analyses stratified by participant sex, yielding nearly identical results (Supplement, Section 2.3).

4. Estimating the category boundary location

4.1. Statistical methods

We next estimated the location of the category boundary, defined as the MH score at which the proportion of participants categorizing the face as "human" is closest to 50%. To do so, we used unweighted ordinary least squares regression to model the proportion of participants categorizing each face as "human" as a polynomial function of MH score, again choosing the best-fitting and most parsimonious polynomial using the AIC. For this analysis, we made a post hoc decision to exclude the 54 faces (30%) that were never categorized as "human" because this large mass of faces with a 0% probability would have been challenging to fit accurately using a smooth polynomial model, and regardless, faces so distant from the category boundary would have contributed little to statistically estimating the boundary location[3]. (No faces had a 100% probability of being categorized as human.) We used the estimated coefficients from this model to estimate the category boundary location.

4.2. Results

Fig. 5 shows the best-fitting model for the relationship between a face's MH score and its probability of being categorized as "human". As expected, the estimated probability of a face's being categorized as "human" increased monotonically with increasing MH score. We estimated that the category boundary occurred at an MH score of 42.5.

[2] For statistical efficiency, we had planned to analyze data at the individual trial level rather than aggregating by face. However, a comparison of individual-level versus face-level estimates of the Uncanny Valley (Supplement, Section 2.3) suggested substantial attenuation of the relationship between MH score and likability in the individual-level data. This strongly suggested that participants' individual ratings of MH score are essentially noisy measurements of a face's "true" MH score and hence that conducting analyses at the individual level would result in downward-biased estimates of the relationship between MH score and likability, a phenomenon that is well-characterized in the literature on nondifferential exposure measurement error (Thomas, Stram, & Dwyer, 1993). This bias is largely mitigated through aggregation as presented in the main text (Prentice & Sheppard, 1995).

[3] A sensitivity analysis in which we did not exclude these faces yielded a very similar estimate of the boundary location.

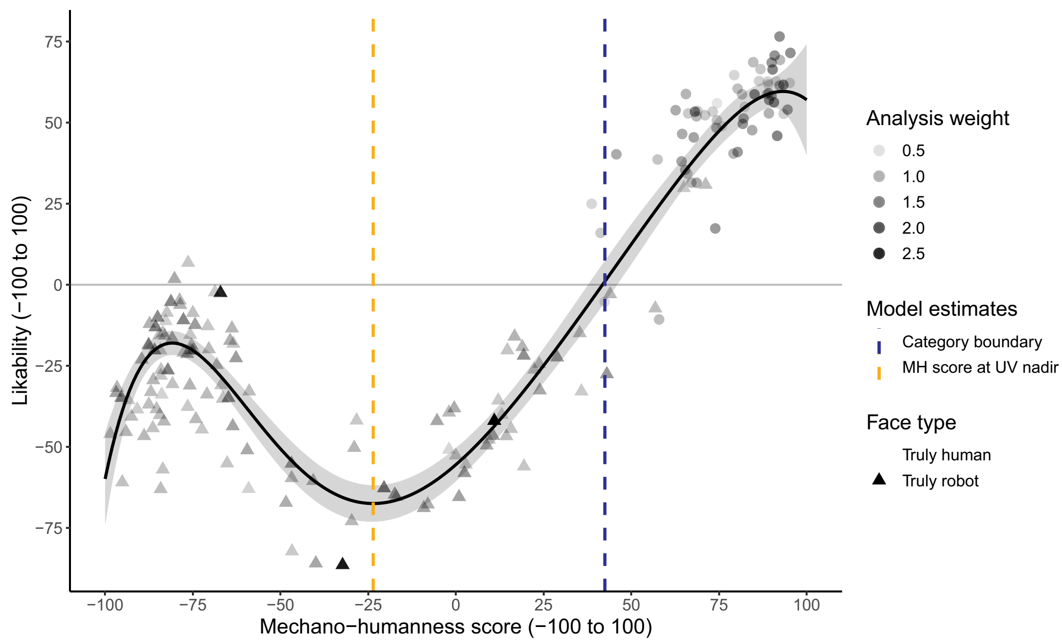

Fig. 4. Uncanny Valley curve in 182 faces (points) with empirically estimated curve based on polynomial regression (solid black curve). Triangles indicate faces that were actually robots; circles indicate faces that were actually humans. Point opacity is proportional to the face's inverse-variance weight in analysis. The shaded band is a 95% pointwise confidence interval for the fitted likability values when setting emotion (the adjusted covariate) to its mean. The vertical dashed lines mark the estimated MH scores at which likability reached its lowest point (orange) and where the category boundary occurred (blue; estimation described in Section 4). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

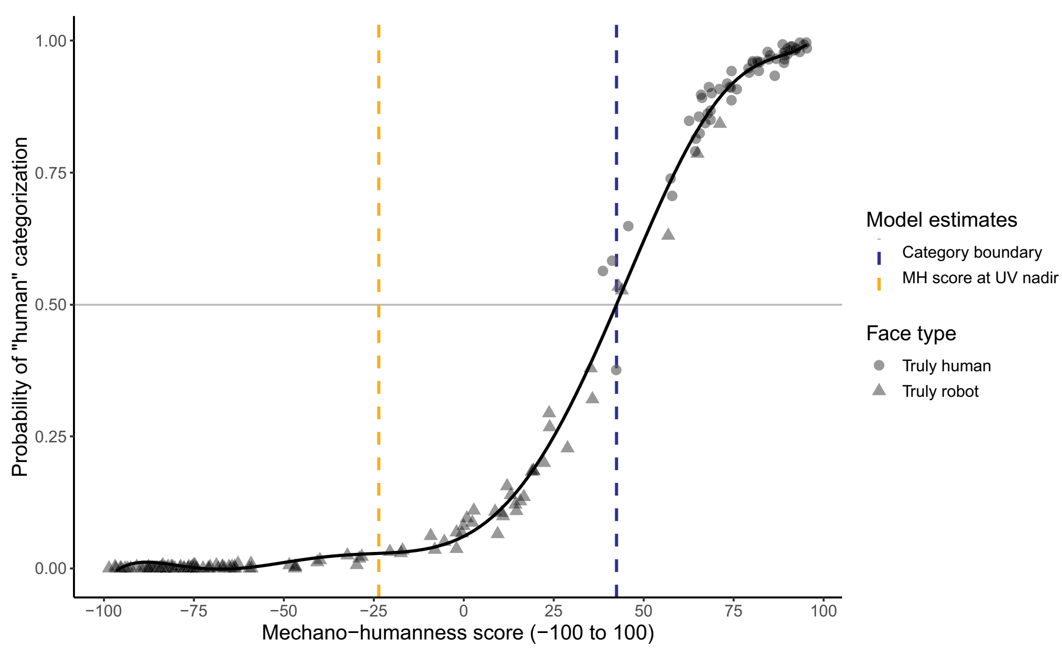

Fig. 5. MH score versus proportion of participants categorizing the face as “human” with empirically estimated curve based on polynomial regression (solid black curve). Triangles indicate faces that were actually robots; circles indicate faces that were actually humans. The vertical dashed lines mark the estimated MH scores at which likability reached its lowest point (orange) and where the category boundary occurred (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

5. Assessing category confusion as a mediator

5.1. Statistical methods

To investigate whether category confusion might be a mechanism of the observed Uncanny Valley effect, we conducted separate mediation analyses for each of the primary and secondary measures of category confusion, treating MH score as the exposure, the confusion measures as the mediators, and likability as the outcome. To improve the comparability of the direct and indirect effect estimates, we standardized MH score and all mediators for these analyses (i.e., mean-centered them and divided by their standard deviations). In addition, as a simple method to consider in aggregate the three mediators pre-specified as primary measures of confusion (i.e., x-flips, maximum x-deviation, and area), we conducted a final mediation analysis on a composite measure of confusion constructed by summing the z-scores for these three primary confusion measures, which were very highly correlated[4] (Pearson's r from 0.91 to 0.99).

Estimating causal mediation effects relies on certain no-confounding assumptions (e.g., VanderWeele (2015)). To this end, during stimulus validation, we had attempted to eliminate many sources of confounding by selecting stimuli that were comparable on graphical features and that were almost emotionally neutral; additionally, in analysis, we controlled for a face's mean emotion as rated during stimulus validation. We conducted mediation analyses using a simulation-based method that involves fitting a model for the mediator as a function of the exposure and a model for the outcome as a function of the mediator and the exposure (Imai, Keele, Tingley, & Yamamoto, 2011; Tingley, Yamamoto, Hirose, Keele, & Imai, 2014). For each mediator model, we used generalized additive models (GAM) with the identity link to regress the measure of category confusion on a spline basis for MH score. For the outcome models, we similarly used GAM to model likability as a function of each mediator and MH score. The outcome models additionally allowed for nonlinear interactions between MH score and the candidate mediator via a tensor product term, which we dropped from the model if its inclusion worsened the model's AIC. We chose these models in order to flexibly accommodate the expected nonlinearities that characterize the Uncanny Valley, as well as the possibly interactive relationship between MH score and confusion. Because the category confusion measures were often skewed or bimodal, suggesting non-normal errors, we estimated all confidence intervals and p-values using nonparametric bootstrapping.

5.2. Results

Figs. 6 and 7 respectively plot MH score versus each confusion measure, and each confusion measure versus likability, along with GAM fits for each relationship. MH score appeared to have nonlinear and non-monotonic relationships with most of the confusion measures; the GAM models for the three primary mediators and their composite estimated that the confusion measures peaked at unstandardized MH scores of 34.3 for x-flips, 95.3 for area, 95.3 for maximum x-deviation, and 95.3 for their composite (Fig. 6, dashed vertical lines). Results were similar for the two secondary confusion measures, namely speed and reaction time. Considering the outcome models, likability did not appear to increase monotonically as the confusion measures increased; rather, the relationships appeared nonlinear and variable across confusion measures (Fig. 7).

Table 1 presents results of the mediation analyses, in which coefficient estimates represent estimated differences in likability on its original scale ranging from −100 to +100. The direct effects represent effects of MH score on likability that occurred independently of each mediating confusion measure; these ranged from 30.0 (95% CI: [12.6, 39.5]) to 32.7 (95% CI: [19.7, 44.9]) for the primary raw measures (x-flips, maximum x-deviation, and area) and 30.0 (95% CI: [8.9, 45.0]) for the composite measure. The indirect effects represent effects of MH score on likability occurring because of mediation by each confusion measure, and the percent mediated represents the percent of the total effect of MH score on likability that is due to mediation by each confusion measure. The estimated indirect effects were approximately an order of magnitude smaller than the estimated direct effects, ranging from −0.6 (95% CI: [-3.6, 2.1]; estimated percent mediated: −2%) to 2.8 (95% CI: [-1.9, 15.4]; estimated percent mediated: 8%) for the three primary confusion measures and 2.2 (95% CI: [-5.8, 20.7]; estimated percent mediated: 7%) for their composite. (Note that negative estimates for the percent mediated occur when the direct and indirect effects are in different directions.) Results for the two secondary confusion measures were qualitatively similar.

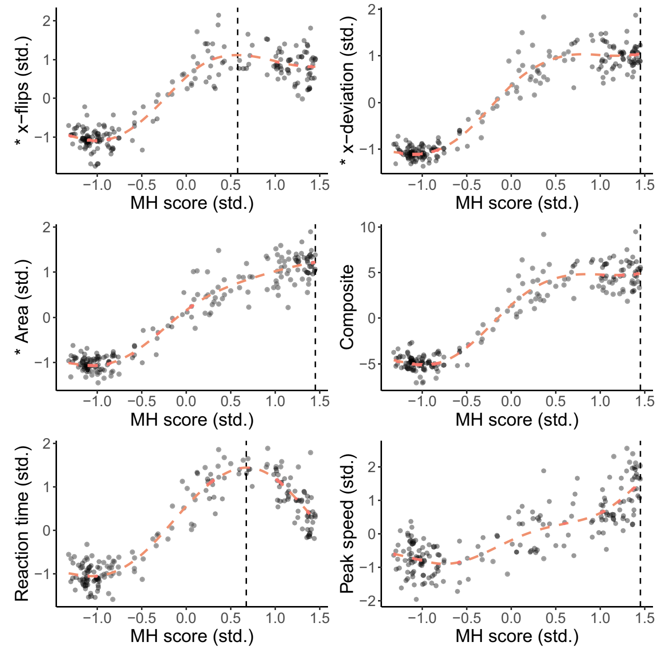

Fig. 6. MH score vs. confusion relationships. The red dashed line represents fitted values from a GAM model as used as in mediation analysis. The vertical dashed line marks the MH score associated with maximum confusion, as estimated by GAM. *: primary confusion measure. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

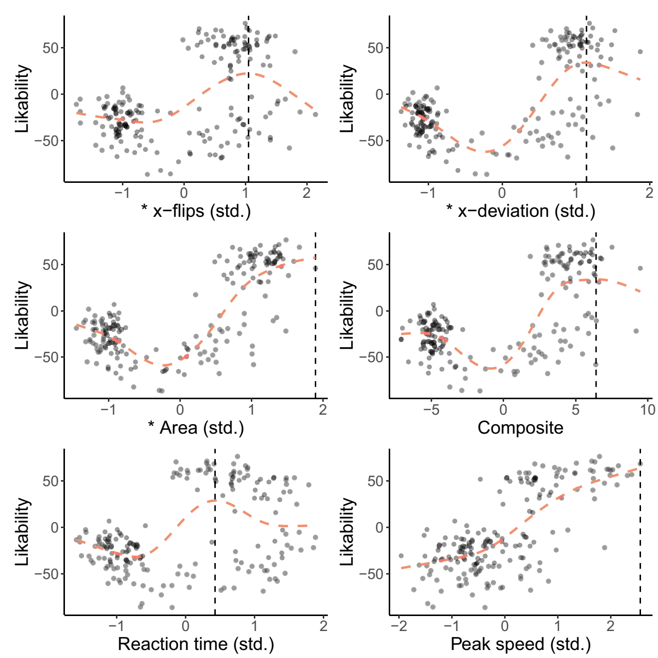

Fig. 7. Confusion vs. likability relationships. The red dashed line represents fitted values from a GAM model without interactions between MH score and mediators. The vertical dashed line marks the value of the confusion measure associated with maximum likability, as estimated by GAM. *: primary confusion measure. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

In summary, the confusion measures varied somewhat in their relationships with MH score. For two of the three primary confusion measures, as well as their composite, faces nearly identical to humans

[4] Note also that analyses handling multiple mediators separately essentially assume that the mediators do not affect one another in a causal sequence, which seems reasonable here given that we measured them at the same time (VanderWeele, 2015). However, it is conceivable that the confusion measures in fact did affect one another in extremely rapid succession; this possibility was another motivation for aggregating the mediators into a composite measure.

Table 1
Analysis of category confusion measures as mediators of the relationship between MH score and likability. Direct and indirect effects are presented for a contrast representing an increase in MH score from its mean to half a standard deviation above its mean. All mediators and MH score were standardized.

Confusion variable	Statistic	Estimate [95% CI]	p-value
Primary measures			
x-flips	Direct effect	32.7 [19.7, 44.9]	< 0.00001
	Indirect effect	−0.6 [−3.6, 2.1]	0.68
	% mediated	−2 [−12, 7]	0.68
x-deviation	Direct effect	31.3 [15.8, 44.7]	0.006
	Indirect effect	0.9 [−5.9, 13.0]	0.62
	% mediated	3 [−22, 41]	0.62
Area	Direct effect	30 [12.6, 39.5]	0.004
	Indirect effect	2.8 [−1.9, 15.4]	0.17
	% mediated	8 [−6, 48]	0.17
Composite	Direct effect	30 [8.9, 45.0]	0.03
	Indirect effect	2.2 [−5.8, 20.7]	0.33
	% mediated	7 [−19, 65]	0.33
Secondary measures			
Reaction time	Direct effect	36.7 [16.3, 49.2]	0.004
	Indirect effect	−5 [−12.1, 10.0]	0.49
	% mediated	−16 [−53, 33]	0.49
Peak speed	Direct effect	29.1 [15.4, 43.1]	0.002
	Indirect effect	3.4 [−3.3, 10.6]	0.13
	% mediated	10 [−11, 34]	0.13

(MH scores near +100) produced the most confusion. For the final primary confusion measure, faces at a lower MH score of 34.3 produced the most confusion. Despite these variations across confusion measures, all indicated that confusion peaked for robots that were considerably more human-like than those that were most dislikable (i.e., those occupying the lowest point of the Uncanny Valley, estimated to occur at an MH score of −23.6). The relationships between the confusion measures and likability suggested that increased confusion was not clearly and monotonically associated with decreased likability, as the category confusion hypothesis might have predicted. Indeed, the mediation analyses suggested that any mediation by confusion was likely very minimal, with results quite consistent across confusion measures.

6. Discussion

The first two aims of our study were to precisely estimate the shape of the Uncanny Valley curve and the location of the boundary between the categories "robot" and "human". To this end, we developed and validated a publicly available corpus of 182 images of real, socially interactive robots as well as humans. These closely controlled faces showed all key features of the theorized Uncanny Valley: an initial increase in likability as faces progressed from extremely mechanical to somewhat less mechanical, followed by a classic "Uncanny Valley" low point as faces became considerably more human-like and markedly dislikable, followed by a gradual increase in likability to its eventual maximum as faces became nearly indistinguishable from humans.

Our estimated Uncanny Valley curve also differed from traditional predictions. The initial apex of likability for unmistakably mechanical faces was negative (likability = −18.0; Fig. 4) rather than markedly positive as Mori (1970) originally postulated (c.f. Fig. 1), indicating that these robots were still somewhat disliked, and very few individual faces in this region did achieve positive likability ratings. This suggests that attempting to avoid negative reactions by deliberately designing android robots that are unmistakably mechanical (Duffy, 2003; Mori, 1970) might severely stunt the robots' social success. Additionally, the lowest point of the Uncanny Valley occurred not for faces nearly indistinguishable from humans, as Mori (1970) originally theorized, but rather for faces perceived to be still predominantly mechanical (i.e., MH score = −23.6). These findings may help direct future research toward exploring hypothesized mechanisms of the Uncanny Valley that are consistent with this precisely estimated shape and in particular its interesting deviations from the usual theoretical predictions.

In the categorization task, the relationship between the faces' human-likeness and their probabilities of being categorized as "human" suggested that humans do perceive a perceptual categorical boundary, as opposed to a smooth continuum, between the properties of "robot" and "human". That is, typical of categorical perception (e.g., de Gelder, Teunisse, and Benson (1997); Etcoff and Magee (1992)), faces in most regions of the human-likeness spectrum were reliably classified as either "robot" or "human", but faces within a steeply sloped region around the category boundary elicited much less stable categorizations (Fig. 4). This boundary zone is where category confusion is thought to occur. It is interesting that our estimated category boundary occurred at an MH score of 42.5, corresponding to faces perceived to be about 71% human on the 200-point MH scale, rather than 50% as one might predict by analogy with classic psychophysical studies of category confusion (Harnad, 1987). Conceivably, this off-center position of the category boundary might relate to the fact that faces' positions on the MH scale were necessarily determined based on participants' subjective ratings rather than objective metrics, since the faces represented real robots rather than constructed stimuli. However, casting some doubt on this interpretation, previous studies using morphing methods to generate objectively quantified mixtures of robot and human faces reported similarly located category boundaries in the range of 60–70% human (Cheetham, Suter, & Jäncke, 2011; Weis & Wiese, 2017).

Critically, the Uncanny Valley and the estimated category boundary (at MH score = 42.5) did not coincide, instead occurring in distinct regions of the human-likeness spectrum. Maximally dislikable faces (MH score = −23.6) were almost always categorized as "robots"; in contrast, unstable categorization occurred primarily at higher MH scores from approximately 0 to 75. Conversely, maximally ambiguous faces were not, on average, disliked. This discrepancy casts doubt on the category confusion hypothesis, which would predict that the most ambiguous faces would be most disliked. For the third aim of our study, we assessed category confusion as a possible mechanism for Uncanny Valley effects by conducting mediation analyses treating human-likeness as the exposure, confusion as the mediator, and likability as the outcome. These analyses did not support mediation by confusion, with indirect effects typically an order of magnitude weaker than direct effects.

These analyses have some limitations. Ideally, we would have measured perceptions of human-likeness, confusion, and likability as they occurred in real time, within a span of milliseconds. This hypothetical design would have clarified, for example, the direction of causation between perceptions of human-likeness and likability. However, such a design would be logistically infeasible, requiring participants to perform three different rating tasks within milliseconds. Furthermore, designs in which participants rate all three characteristics in quick succession may introduce task interference or demand characteristics. Our three-wave design was intended to minimize these types of bias, but cannot rule out reverse causation. Additionally, as discussed, all mediation analyses rely on no-confounding assumptions. We tried to minimize confounding by selecting closely matched stimuli and controlling for perceived emotion in analyses. Given the very small size of the indirect effects, it seems unlikely that residual confounding would have masked meaningfully strong mediation effects.

These findings suggest that category confusion is not a viable explanation for the Uncanny Valley. One promising alternative hypothesis, called "feature inconsistency" or "perceptual mismatch", postulates instead that inconsistent realism across features in a face (for example, mechanical-looking eyes and mouth on an otherwise human-like face) drives Uncanny Valley reactions (MacDorman & Chattopadhyay, 2016). This hypothesis is subtly, but importantly, distinct from category confusion. For example, two faces occupying the same position relative to the category boundary (i.e., having the same MH score) could nevertheless have different amounts of feature inconsistency; the more consistent face might have all moderately human-like features, whereas the less consistent face might have some highly mechanical and some highly human-like features. Future work could use rigorous statistical analyses and well-controlled, realistic stimuli, such as the public corpus from the present study, to assess whether feature inconsistency is associated with Uncanny Valley reactions independent of MH score, and whether inconsistency is a mediator of the relationship between MH score and likability.

To elucidate the mechanisms and boundary conditions of the Uncanny Valley, we would also suggest holistic consideration of negative reactions to mid-range stimuli in general, including existing work outside the immediate scope of the Uncanny Valley literature. For example, heuristically, patterns in the categorization and social judgment of interracial human faces (Freeman et al., 2016) seem to resemble the Uncanny Valley in android robots. Might such effects share fundamental mechanisms with the Uncanny Valley? A key step will be to examine which potentially related effects share the distinctive subjective quality of Uncanny Valley reactions, a farrago variously described as "eerieness", "creepiness", and "repulsiveness" (Ho & MacDorman, 2010; Moore, 2012). Suggestive evidence for or against proposed mechanisms might also arise from continued investigation of whether the Uncanny Valley is experienced universally (Koopman, 2019) and investigation of individual-level moderators of sensitivity to its effects, such as an individual's frequency of exposure to android robots, attitudes toward robots, and affective states (MacDorman & Entezari, 2015; Łupkowski & Gierszewska, 2019).

Overall, our findings suggest that although humans do perceive a category boundary between "robot" and "human", the location of this boundary does not coincide with the Uncanny Valley itself, and category confusion produced by this boundary does not seem to explain Uncanny Valley aversions. It is striking that, despite the decades-long prominence of the Uncanny Valley theory and robot designers' sophisticated attempts to circumvent it, the robots we sampled — which were purposefully designed for social interaction — nevertheless were dislikable on average (mean likability = −32.9) and showed a prominent Uncanny Valley. These findings point to the continued importance of attempting to elucidate the mechanisms underlying the effect.

Acknowledgments

This research was funded by a grant from the Harvard University Mind, Brain, & Behavior Initiative. The funders had no role in the conduct or reporting of this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2019.08.029.

Reproducibility

All materials, data, and code required to reproduce this research are publicly available and documented (https://osf.io/2v6f4/). The pre-registrations are publicly available (https://osf.io/mu5xj/registrations).

Author contributions

MBM and DBR conceptualized and designed the study. MBM planned and conducted statistical analyses and led writing. The remaining authors collected data and revised the manuscript.

References

Akaike, H. (1974). A new look at the statistical model identification. *Selected papers of hirotugu Akaike* (pp. 215–222). Springer.

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*(4), 651.

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., & Jancke, L. (2013). Category processing and the human likeness dimension of the Uncanny Valley hypothesis: Eye-tracking data. *Frontiers in Psychology, 4*, 108.

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.

Cheetham, M., Suter, P., & Jancke, L. (2014). Perceptual discrimination difficulty and familiarity in the Uncanny Valley: More like a "happy valley". *Frontiers in Psychology, 5*, 1219.

Cheetham, M., Wu, L., Pauli, P., & Jancke, L. (2015). Arousal, valence, and the uncanny valley: Psychophysiological and self-report findings. *Frontiers in Psychology, 6*, 981.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3–4), 177–190.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44*(3), 227–240.

Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: A novel account of the Uncanny Valley. *Frontiers in Psychology, 6*, 249.

Fiske, S. T., & Taylor, S. E. (2008). *Social cognition: From brains to culture.* New York: McGraw Hill.

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology: General, 137*(4), 673.

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences, 20*(5), 362–374.

Freeman, J. B., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed- race bias. *Psychological Science, 27*(4), 502–517.

de Gelder, B., Teunisse, J.-P., & Benson, P. J. (1997). Categorical perception of facial expressions: Categories and their internal structure. *Cognition & Emotion, 11*(1), 1–23.

Harnad, S. (1987). A new look at the statistical model identification. In S. Harnad (Ed.). *Categorical perception: The groundwork of cognition* (pp. 1–25). Cambridge University Press.

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*(6), 1508–1518.

Ho, C.-C., MacDorman, K. F., & Pramono, Z. D. (2008). Human emotion and the Uncanny Valley: A GLM, MDS, and isomap analysis of robot video ratings. *3rd ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 169–176). .

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765–789.

Jentsch, E. (1906). *Zur psychologie des unheimlichen.* Psychiatrisch-Neurologische Wochenschrift.

Jentsch, E. (1997). On the psychology of the uncanny (1906). *Angelaki: Journal of the Theoretical Humanities, 2*(1), 7–16.

Jung, Y., & Cho, E. (2018). Context-specific affective and cognitive responses to humanoid robots. *The 22nd biennial conference of the international telecommunications society: "beyond the boundaries: Challenges for business, policy and society".* Seoul: International Telecommunications Society (ITS).

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*(8), 759.

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*.

Kawabe, T., Sasaki, K., Ihaya, K., & Yamada, Y. (2017). When categorization-based stranger avoidance explains the Uncanny Valley: A comment on MacDorman and Chattopadhyay (2016). *Cognition, 161*, 129–131.

Koopman, R. (2019). *The uncanny valley as a universal experience: A replication study using multilevel modelling*B.S. thesis. University of Twente.

Lay, S., Brace, N., Pike, G., & Pollick, F. (2016). Circling around the uncanny valley: Design principles for research into the relation between human likeness and eeriness. *I-Perception, 7*(6) 2041669516681309.

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.

Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science, 21*(12), 1854–1862.

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

MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies, 16*(2), 141–172.

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

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

Mathur, M. B., & Reichling, D. B. (2019). Open-source software for mouse-tracking in Qualtrics to measure category competition. *Behavior Research Methods.* https://doi.org/10.3758/s13428-019-01258-6.

Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific Reports, 2*, 864.

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

Prentice, R. L., & Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika, 82*(1), 113–125.

Rabbitt, S. M., Kazdin, A. E., & Scassellati, B. (2015). Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. *Clinical Psychology Review, 35*, 35–46.

Schwind, V., Leicht, K., Jäger, S., Wolf, K., & Henze, N. (2018). Is there an Uncanny Valley of virtual animals? A quantitative and qualitative investigation. *International Journal of Human-Computer Studies, 111*, 49–61.

Slijkhuis, P. J. (2017). *The uncanny valley phenomenon: A replication with short presentation times*Unpublished master's thesis. University of Twente.

Thomas, D., Stram, D., & Dwyer, J. (1993). Exposure measurement error: Influence on exposure-disease relationships and methods of correction. *Annual Review of Public Health, 14*(1), 69–93.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). *Mediation: R package for causal mediation analysis*.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction.* Oxford University Press.

Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology, 19*(4), 393.

Weisberger, M. (2018). *A floating "brain" will assist astronauts aboard the space station.* https://www.livescience.com/61909-cimon-floating-brain-iss.html, Accessed date: 20 April 2019.

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: Vol. 61*, (pp. 1599–1603).

Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. *Japanese Psychological Research, 55*(1), 20–32.

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