Cognition 146 (2016) 22–32

Contents lists available at ScienceDirect

Cognition

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

Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley

Maya B. Mathur[a,*], David B. Reichling[b]

[a] Quantitative Sciences Unit, Stanford University, United States
[b] Department of Oral & Maxillofacial Surgery, University of California, San Francisco, United States

ARTICLE INFO

Article history:
Received 4 July 2015
Revised 6 September 2015
Accepted 7 September 2015
Available online 21 September 2015

Keywords:
Human–robot interaction
Social interaction
Game theory
Facial appearance
Trust
Social decision-making

ABSTRACT

Android robots are entering human social life. However, human–robot interactions may be complicated by a hypothetical Uncanny Valley (UV) in which imperfect human-likeness provokes dislike. Previous investigations using unnaturally blended images reported inconsistent UV effects. We demonstrate an UV in subjects' explicit ratings of likability for a large, objectively chosen sample of 80 real-world robot faces and a complementary controlled set of edited faces. An "investment game" showed that the UV penetrated even more deeply to influence subjects' implicit decisions concerning robots' social trustworthiness, and that these fundamental social decisions depend on subtle cues of facial expression that are also used to judge humans. Preliminary evidence suggests category confusion may occur in the UV but does not mediate the likability effect. These findings suggest that while classic elements of human social psychology govern human–robot social interaction, robust UV effects pose a formidable android-specific problem.

1. Introduction

Robots are no longer merely features of our technological environment, but are beginning to penetrate our social sphere (Breazeal, 2003; Fong, Nourbakhsh, & Dautenhahn, 2003; Zhao, 2006), and people who interact with robots are increasingly unlikely to be technically trained experts and thus more likely to use casual intuitive approaches to the interaction. Unexpectedly negative reactions to imperfectly human robots have become a major problem in the design of socially interactive robots. This phenomenon (Fig. 1), termed the "Uncanny Valley" (Mori, 1970), has dominated discussion of human reactions to anthropomorphic robots in both popular culture and research literature. Despite its prominence, the existence of an Uncanny Valley (UV) is controversial (Burleigh, Schoenherr, & Lacroix, 2013; Hanson, 2006; Katsyri, Forger, Makarainen, & Takala, 2015; MacDorman, Green, Ho, & Koch, 2009), and a recent systematic review concluded that "empirical evidence for the uncanny valley hypothesis is still ambiguous if not non-existent" (Katsyri et al., 2015). Most studies attempting to address the issue have employed progressively morphed blends of human and robot faces, in which two face images are digitally overlaid with varying degrees of opacity, in some cases enhanced by warping of features in intermediate images. This method introduces unnatural distortions, such as semi-transparent or bent facial features, that would be most prominent in the more highly processed images in the midrange of a morphed face series. This could potentially create an UV-like artifact in that region (Katsyri et al., 2015).

The present study was designed to determine if human reactions to android robots truly exhibit an UV effect, and if so, to determine the degree to which it actually influences humans' willingness to trust a robot as a social partner. Experiment 1 examined human reactions to a large, objectively chosen sample of real-world android robots using subjects' explicit judgments of the mechano-humanness and likability of each face. Next, to determine whether the influence of a potential UV actually penetrates humans' implicit social decision-making, we employed game-theory methodology to measure subjects' practical inferences (as measured by real financial risk-taking) concerning the trustworthiness of each robot. An exploratory analysis tested the theory that UV effects arise from perceptual category confusion.

In contrast to the large, heterogeneous population of wild-type robots (with variable facial expressions, positions, and background settings) of Experiment 1, Experiment 2 took a complementary approach: we used a precisely controlled series of 6 digitally composed robot faces with constant morphometry to assess social responses to a single face configuration in its controlled progression from mechanical to human. In addition, this control over the

[*] Corresponding author at: Quantitative Sciences Unit, 1070 Arastradero Rd, Palo Alto, CA, United States.

E-mail address: mmathur@stanford.edu (M.B. Mathur).

http://dx.doi.org/10.1016/j.cognition.2015.09.008
0010-0277/© 2015 The Authors. Published by Elsevier B.V.

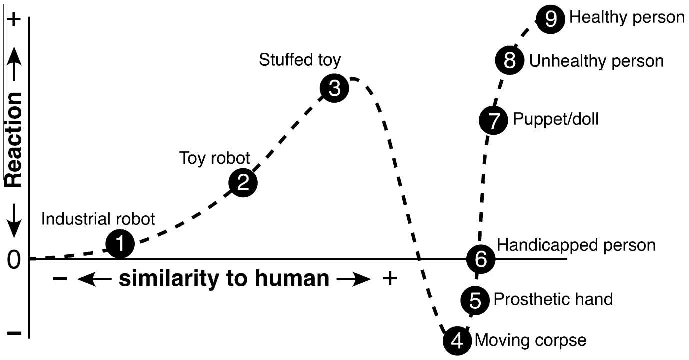

Fig. 1. The Uncanny Valley conjecture, adapted from Mori (1970).

face images allowed us to measure the effect on trust-motivated behavior of a subtle change in facial expression.

2. Experiment 1: wild-type robots

2.1. Experiment 1A: quantifying the mechano-humanness spectrum

2.1.1. Methods

All protocols received IRB approval. Experiments 1A–1C used the same sample of 80 real-world robot faces (Fig. 2) that embodied the myriad design choices made by actual robot designers, choices that may be subtle and unexpected and may vary depending on whether the designer's intention is to build more mechanical versus more human-like robots. The size of the sample and its diversity in mechano-humanness enabled a fine-grained statistical analysis of the effect of mechano-humanness on human social perceptions. To reduce bias in selecting the robots or their manner of presentation (expressions, poses, viewing angles, background settings, etc.), we conducted a systematic search using specific inclusion and exclusion criteria. We performed four Google image searches on a single day using the following sets of search terms: "robot face," "interactive robot," "human robot," and "robot."

Inclusion criteria were:

1. Full face is shown from top of head to chin.
2. Face is shown in frontal to 3/4 aspect (both eyes visible).
3. The robot is intended to interact socially with humans.
4. The robot has actually been built.
5. The robot is capable of physical movement (e.g., not a sculpture or purely CGI representation that lacks a three-dimensional body structure).
6. The robot is shown as it is meant to interact with users (e.g., not missing any hair, facial parts, skin, or clothing, if these are intended).
7. The robot represents an android that is plausibly capable of playing the wagering game (e.g., not a baby or an animal).
8. The resolution of the original image (or an exact copy when one could be located) is sufficient to yield a final cropped image at 100 d.p.i. and 3 in. tall.

Exclusion criteria were:

1. The robot represents a well-known character or a famous person (e.g., Einstein).
2. The image includes other faces or human body parts that would appear in the final cropped image.
3. Objects or text overlap the face.
4. The robot is marketed as a toy.

When the search returned multiple images of a particular robot, we accepted only the first image encountered; if an image failed only graphical criteria, we accepted the next graphically adequate image of the same robot. We accepted the first 80 face images satisfying inclusion criteria and cropped them to include top of head to bottom of chin (or when those features were missing, images were similarly framed in approximate proportion to the features).

For Experiments 1A–1C, we sampled subjects via Amazon Mechanical Turk, a crowdsourcing platform allowing workers to complete brief online tasks in exchange for pay. The task title and description were vague to minimize sampling bias and demand characteristics. We sampled United States workers with excellent performance history (>95% of previous online tasks "approved" as high-quality by requester). By contractual agreement with Amazon, workers must be at least 18 years old. Mechanical Turk workers tend to be somewhat younger, more educated, and lower-income than the US general population, but are demographically more representative than typical university-based research samples (Paolacci, Chandler, & Ipeirotis, 2010). Studies performed on Mechanical Turk can yield high-quality data, minimize experimental biases, and successfully replicate the results of behavioral studies performed on traditional samples (Paolacci et al., 2010). No subjects were duplicated across any of the experiments to avoid effects of previous exposure to the stimuli.

The purpose of Experiment 1A was to determine (1) the degree to which each robot face is perceived as exhibiting human and mechanical properties, (2) the extent to which human-resemblance and mechanical-resemblance behave as a unidimensional property, and (3) secondarily, the perceived valence and magnitude of emotion displayed by each face, which might strongly influence and confound social responses in subsequent experiments (Scharlemann, Eckel, Kacelnik, & Wilson, 2001).

In an online questionnaire, subjects first viewed a page of thumbnails of all 80 face stimuli (similar to Fig. 2 but with faces arbitrarily positioned with respect to mechano-humanness) to give them a sense of the range of the faces they would encounter. Subjects then viewed and rated each of the 80 robot faces one at a time, with the order of faces randomized for each subject. The rating scale was a continuous visual analog scale (VAS) without graduations, which can provide more precise and psychometrically valid ratings than a Likert-type ordinal scale (Reips & Funke, 2008). The subjects controlled how long they viewed each face with no time limit. If individual subjects had rated both the mechanical- and human-resemblance of a face, they might have assumed some relationship between the two properties to which their ratings should conform (e.g., sum to 100). Therefore, in

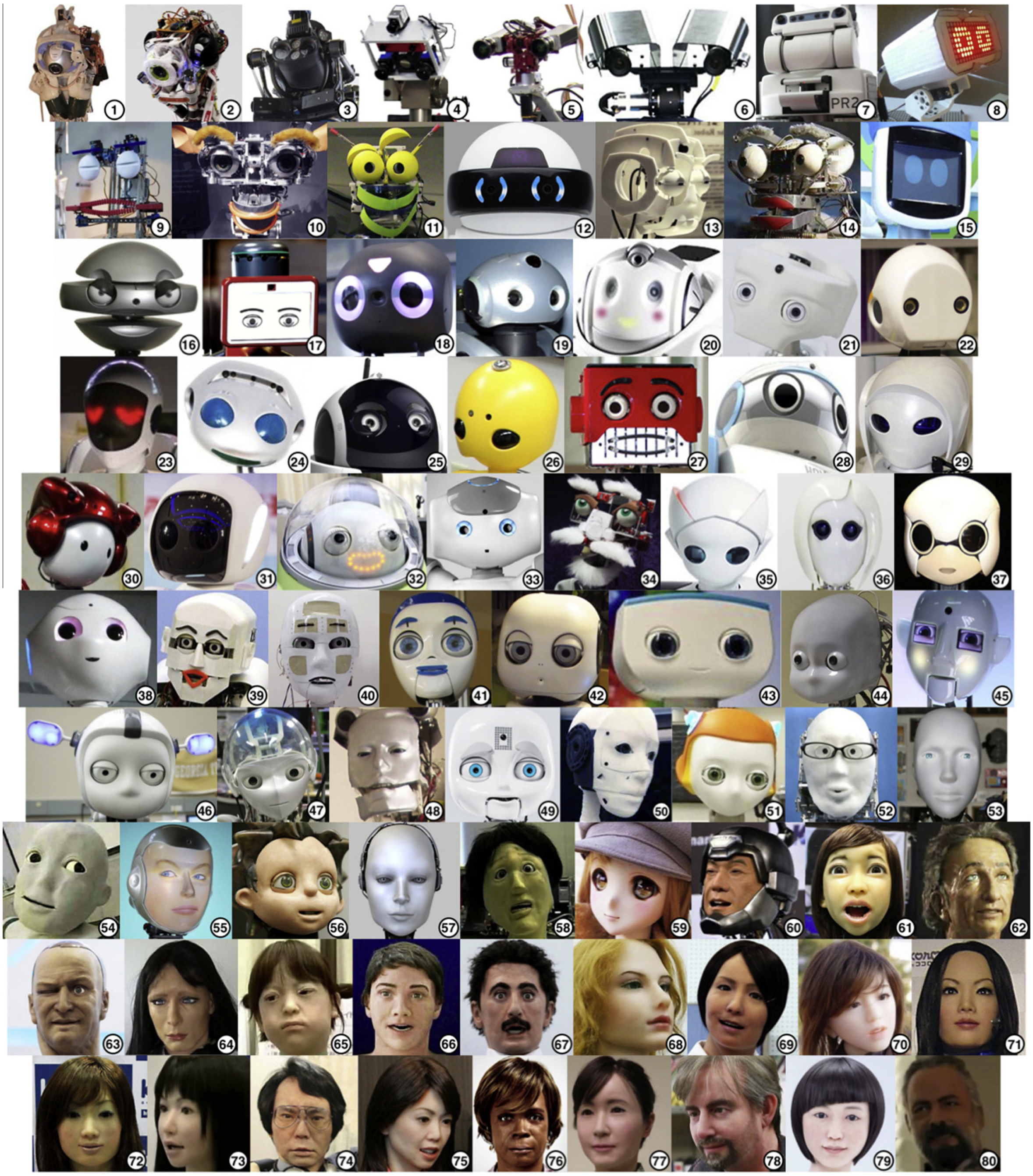

Fig. 2. Wild-type robot face stimuli (Experiment 1A–C) numbered and displayed in ascending order of mechano-humanness score estimated in Experiment 1A. Face #54 was rated closest to the scale's midpoint (score −4.5).

contrast to the approach used in a previous study (Rosenthal-von der Pütten & Krämer, 2014), subjects were randomized to rate all the faces they would be shown according to only one of the two properties, being asked either "How mechanical does this robot face look?" (0–100 scale) or "How human does this robot face look?" (0–100 scale). Additionally, every subject answered the question, "How much positive or negative emotion is this robot face showing?" (−100 to +100 scale, where −100 represented strongest negative emotion and +100 represented strongest positive emotion). To detect subjects who might have rushed excessively through the large number of face stimuli, we included an attention-check consisting of a robot face (not otherwise used in the survey) with mouth superimposed by red letters directing subjects to rate that face exactly 41 on the VAS.

We performed all analyses in R, Version 3.0.2,[1] defined statistical significance at $\alpha = 0.05$, and used two-tailed tests. We treated face stimulus as the unit of analysis, descriptively characterizing each face by its mean human-resemblance, mechanical-resemblance, and perceived emotion. We used the Pearson correlation to assess the relationship between human- and mechanical-resemblance. Finally, we characterized coherence of ratings within subjects using the intraclass correlation (ICC), which can be interpreted within an ANOVA framework as the proportion of total

[1] We used the following R packages: *car, lme4, lmerTest, reshape2, Hmisc, plotly, lessR, lsmeans, ggplot2, data.table, lessR, lavaan.*

variance due to across-subject variability. A higher ICC indicates more clustering of responses within subjects.

2.1.2. Results

Subjects ($n = 66$)[2] were 56% female, majority Caucasian (85%), and had median age 32 years. We excluded 78 additional subjects who failed the stringent attention check.[3] The subjects spent a median of 12.2 s to rate each face. Treating face as the unit of analysis, face ratings spanned nearly the entire possible range of both human-resemblance (mean = 42, SD = 31, min = 2, max = 97) and mechanical-resemblance (mean = 66, SD = 31, min = 3, max = 99), enabling subsequent experiments to precisely estimate a potential UV in fine detail throughout the entire mechano-humanness range. Human-resemblance and mechanical-resemblance were nearly perfectly correlated (Appendix Fig. A.1; $r = -0.97$; $p < 0.001$). Therefore, in subsequent experiments, we combined the two measures into a single unidimensional scale (the "mechano-humanness" or "MH" score) formed by subtracting mean mechanical-resemblance from mean human-resemblance. The nearly perfect correlation between mechanical-resemblance and human-resemblance of the robot faces suggests that the sample did in fact represent a spectrum of android robots. That is, if the sample had been contaminated with robots perceived as portraying non-human beings (e.g. animals, aliens, mythological creatures), such robots would likely score low on human-resemblance regardless of mechanical-resemblance and thus disrupt the correlation between these dimensions. Ratings of both characteristics showed little clustering within subjects (human-resemblance: ICC = 0.05; 95% CI: 0.03, 0.10; mechanical-resemblance: ICC = 0.05; 95% CI: 0.03, 0.10).

Perceived emotion was uncorrelated with both human-resemblance ($r = 0.006$; 95% CI: −0.21, 0.23; $p > 0.25$) and mechanical-resemblance ($r = -0.04$; 95% CI: −0.26, 0.18; $p > 0.25$), suggesting that the variety of emotional facial expressions encountered in the wild-type robot face sample would not statistically confound the analysis of social responses to the robot faces in subsequent experiments.

2.2. Experiment 1B: perceptions of likability

2.2.1. Methods

To measure the perceived likability of each face, we asked subjects to "estimate how friendly and enjoyable (versus creepy) it might be to interact with each face in an everyday situation" using a VAS ranging from −100 ("Less friendly; more unpleasant and creepy") to +100 ("More friendly and pleasant; less creepy"). For Experiments 1B–1C, subjects viewed and rated a randomly selected subset of 15 of the 80 faces, presented in randomized order for each subject.

We quantitatively modeled the relationship between mean MH score and mean likability (as measured in Experiment 1A, with a higher score indicating increasing human-resemblance and decreasing mechanical-resemblance) via polynomial regression with face stimulus as the unit of analysis ($n = 80$).[4] We used inverse-variance weighting to allow faces with more precise estimates to be weighted more strongly in model fitting. Throughout Experiment 1, we used F-tests to compare the fit of linear models with second-degree, third-degree, and fourth-degree polynomial terms of MH score in order to select the best-fitting and most parsimonious model.

Because we expected a face's perceived emotion to be strongly related to its likability, we additionally refit analysis models among only "low-emotion" faces, defined *a priori* as those estimated in Experiment 1A to occupy the most neutral 10% of the emotion scale (between 10 and +10 on the −100 to +100 scale). This category comprised 50 (63%) of the 80 faces. Additionally, we formally assessed a possible interaction effect of emotion (first treated as continuous, then as a dichotomy between low-emotion and other faces[5]) via an F-test comparing the fit of the main analysis model (plus a main effect of emotion) to the same model plus interactions of perceived emotion with all polynomial terms corresponding to MH score. The latter flexibly allows for interaction effects of emotion on the shape of the UV curve. The first interaction analysis (treating emotion as continuous) addresses whether Uncanny Valley effects change as robots progress along a bipolar, valenced spectrum of emotion; in contrast, the dichotomized interaction test assesses whether the Uncanny Valley effect differs between low-emotion faces and faces with higher emotion (treating all higher-emotion faces as comparable regardless of positive or negative emotional valence).

2.2.2. Results

Subjects ($n = 342$ after the exclusion of 3 subjects reporting substantial technical or comprehension problems) were 40% female,[6] had median age 30 years, and were 80% Caucasian. Because each subject viewed a randomized subset of the face stimuli, each face was rated by 64 subjects on average. Subjects spent a median of 8.1 s responding to each face. Ratings of likability showed little within-subject clustering (ICC = 0.08; 95% CI: 0.06, 0.11), suggesting that individual subjects did not differ greatly in overall propensity to like robot faces in general.

The third-degree model representing the relationship between mechano-humanness and likability – the lowest-degree polynomial able to represent the two inflection points that crucially define Mori's UV (Mori, 1970) – fit significantly better than the lower- and higher-degree models (third- versus second-degree: $F(1) = 20.49$; $p < 0.001$; fourth- versus third-degree: $F(1) = 0.35$; $p > 0.25$).[7] This quantitatively optimized curve ($R^2_{adj} = 0.29$; Fig. 3A, solid line) demonstrates several features central to Mori's conceptualization of the UV. As faces progressed from completely mechanical

[2] For all experiments, we determined sample sizes in advance and ceased data collection after reaching or exceeding the intended minimum sample size targets. Data were kept closed to analysis until the end of data collection, and no subjects were enrolled after analysis began.

[3] After data collection, it became obvious that the attention-check question was unintentionally stringent. Because success on this question was an *a priori* inclusion criterion, we applied this criterion in the main analyses. However, sensitivity analyses in which we did not exclude responses on this basis yielded nearly identical results. Subsequent experiments did not include the attention-check question.

[4] We treated face stimulus as the unit of analysis because entirely separate subject groups rated faces for mechanical- or human-resemblance and for likability – a decision made in order to minimize the influence of demand characteristics with regard to the relatedness of these two properties. Thus, it was not appropriate to treat single observations as the unit of analysis (e.g., via a mixed-effects approach modeling likability or trust by a main effect of MH score and a random intercept by subject). As a sensitivity analysis to assess the impact of accounting for subject-specific effects, we fit a linear mixed-effects regression modeling each outcome by a fixed effect of face stimulus and a random subject intercept. The fitted values of this model for each face can be considered to represent a face-specific mean for likability and trust, respectively, adjusting for subject-specific effects. Characterizing faces by these "adjusted" means rather than simple means yielded nearly identical results to main analyses.

[5] The continuous interaction test was chosen *a priori*, while the dichotomized interaction test was chosen post hoc to clarify results.

[6] There was a significantly lower proportion of female subjects in Experiments 1B–1C than in Experiment 1A. However, because ratings of human- and mechanical-resemblance did not differ by sex in Experiment 1A ($p = 0.70$), it was reasonable to generalize the results of 1A to characterize face means in subsequent experiments.

[7] Inverse-variance weights were symmetrically distributed with mean 0.005 and standard deviation 0.0002.

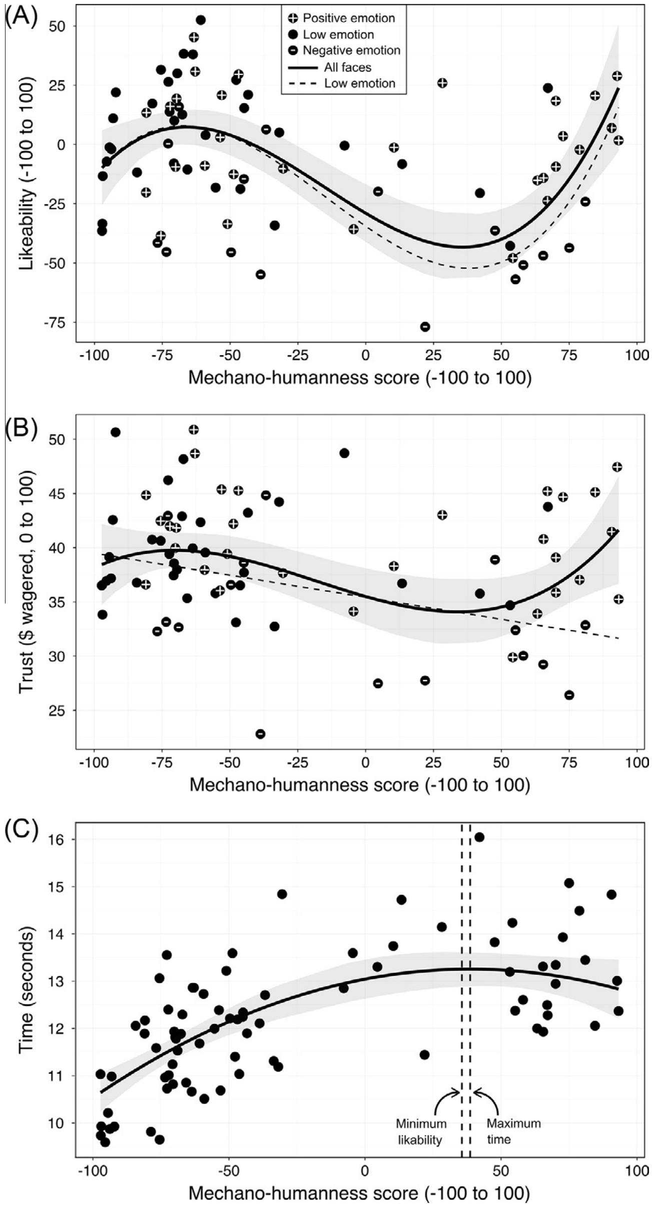

Fig. 3. Uncanny Valley in wild-type robot faces. Fitted curves are shown for likability (panel A; Experiment 1B) and trust-motivated behavior (panel B; Experiment 1C). Solid curves represent best-fitting polynomial regression models and shaded regions show 95% confidence intervals. Dashed curves represent models fit to data from only the 50 low-emotion faces. Panel C shows median rating times in Experiment 1A as a function of MH score; vertical lines mark MH scores associated with maximum rating time and minimum likability.

to completely human-like, likability increased to an initial apex (of +7, approximately neutral) for a moderately mechanical face (MH score −66). At this point, likability began to decline with increasing human-likeness, reaching a fitted nadir of −43 – well below neutral – for a somewhat human-like face (MH score +36). Further increases in human-likeness once again became associated with increased likability until a face that is fully human (MH score +100) had an estimated likability of +43.

As expected, faces perceived as showing more positive emotion were considered more likable ($r = 0.57$; 95% CI: 0.40, 0.70; $p < 0.001$), but because perceived emotion was unassociated with human-likeness (Experiment 1A), emotion could not have confounded a potential UV effect. Although not a statistical confounder, variability in the valence and magnitude of perceived emotion of the robot faces could have acted as a moderator or added substantial noise to this analysis. However, when we conducted the main analysis among only low-emotion faces, results were nearly identical to those obtained with the full set of faces (Fig. 3A, dashed line; third- versus second-degree model: $F(1) = 8.97$; $p = 0.004$; fourth- versus third-degree: $F(1) = 0.49$; $p > 0.25$). Perceived emotionality of the robots did not moderate the relationship between MH score and likability when treated as continuous ($F(3) = 0.08$; $p > 0.25$) or as a dichotomy between low-emotion and higher-emotion faces ($F(3) = 1.23$; $p > 0.25$).

2.3. Experiment 1C: trust-motivated behavior

2.3.1. Methods

This experiment measured subjects' inferences concerning the trustworthiness of each robot face. Rather than ask subjects explicitly to rate the face stimuli on a scale of trustworthiness, we used methods from the field of game theory to attempt to more directly measure their actual willingness to trust the robots in a game with real financial consequences. The subjects engaged in a simple wagering game in which they chose how much money to entrust to each robot. The game was a variant of the classic "investment game" (Berg, Dickhaut, & McCabe, 1995), in which Player A (in this case, the human) decides how much of an endowment of money to pass to Player B (the pictured robot). Player B then decides what portion of the passed amount, if any, will be returned (after having been tripled by the experimenter) to Player A.

Subjects were told prior to beginning the game that their wagers (between $0 and $100) would be "transmitted to the robot laboratories, and the imaginary money will be distributed according to the robots' decisions." Because we were interested only in subjects' initial judgments of trustworthiness based purely on robots' facial appearance, our game did not progress past the first wager. The game was therefore a true one-shot game involving no expectation for future play which could have caused subjects to adopt complicated reputation-building strategies (Mailath & Samuelson, 2006). To encourage thoughtful wagering, subjects were told, "If you are among the 50% best players of the game, you will receive a bonus of $1.00" (which was 50% more than their base pay for the task).

Similarly to Experiment 1B, we modeled each face's perceived trustworthiness (defined as mean dollars wagered) based on its MH score estimated in Experiment 1A and further analyzed as in Experiment 1B.

2.3.2. Results

Subjects ($n = 334$ after the exclusion of 7 subjects reporting technical or comprehension problems) were 42% female, majority Caucasian (75%), and had median age 30 years. Each face stimulus was viewed on average by 63 subjects. Subjects spent a median of 7.5 s responding to each face. Ratings showed moderately strong within-subject clustering (ICC = 0.60; 95% CI: 0.56, 0.64), indicating that some subjects were generally more or less trusting than others.

As in Experiment 1B, the third-degree model for the relationship between MH score and trust was the best-fitting (third- versus second-degree: $F(1) = 4.80$; $p = 0.03$; fourth- versus third-degree: $F(1) = 0.04$; $p > 0.25$).[8] The fitted curve ($R^2_{adj} = 0.07$; Fig. 3B, solid line) suggested that, similarly to explicit reports of perceived likability, trust-motivated behavior follows an "Uncanny Valley" pattern. Faces achieved an initial apex of trustworthiness (earning a predicted wager of $40) when they were predominantly

[8] Inverse-variance weights were left-skewed and bimodal with mean and median 0.001 and standard deviation 0.0002.

mechanical in appearance (MH score –70), declining to a nadir (predicted wager $34) when they were somewhat human-like (MH score +34) and rebounding again to a maximum predicted wager of $44 when they became fully human-like.

As observed for likability, robots perceived as showing more positive emotion tended to elicit more trust ($r = 0.58$; 95% CI: 0.41, 0.71; $p < 0.001$). In contrast to likability, the UV effect was entirely dampened when only the 50 faces displaying low emotion were analyzed; a linear model without inflection points now fit as well as second- or higher-degree polynomials (Fig. 3B, dashed line; second-degree versus linear model: $F(1) = 0.06$; $p > 0.25$). Perceived emotion did not, however, significantly moderate the relationship between MH score and trust when treated as continuous ($F(3) = 0.43$; $p > 0.25$) or when dichotomized ($F(3) = 0.46$; $p > 0.25$).[9]

2.4. Sensitivity analysis excluding the most human-like robots

2.4.1. Methods

To investigate whether the most human-like robots drove the entire observed effect of mechano-humanness, we conducted a post hoc sensitivity analysis in which we removed the top 20 most human-like faces (25%), yielding an analysis sample of the least human-like 60 faces. We used an *F*-test to compare the fit of a third-degree model (as in the main analysis) to that of a null model including only an intercept (but no coefficients for MH score). If the most human-like faces dominated the observed Uncanny Valley effect, the third-degree model would be expected to fit no better than the null model among this restricted subset of faces.

2.4.2. Results

For likability, the third-degree model continued to outperform the null model ($F(3) = 5.19$; $p = 0.003$). For trust, restriction to this subset of faces did reduce the performance of the third-degree model to match that of the null model ($F(3) = 1.57$; $p = 0.20$). Overall, these results suggest that the Uncanny Valley for likability is robust even with a dramatic range restriction to only the most mechanical 75% of faces. The contrasting result for trust is consistent with the most human-like faces tending to drive the Uncanny Valley for trust. On the other hand, trust showed a less dramatic effect than likability in the primary analysis, and would therefore also be more strongly impacted by a loss of power due to a 25% reduction in sample size.

2.5. Exploratory analyses of category confusion

A prominent hypothesis (Katsyri et al., 2015) postulates that the UV arises from ambiguity that is experienced at the boundary between perceptual categories (de Gelder, Teunisse, & Benson, 1997; Repp, 1984) – in this case, between non-human and human categories. Such category confusion is measured experimentally as an increase in the time required to categorize a stimulus (de Gelder et al., 1997; Pisoni & Tash, 1974; Yamada, Kawabe, & Ihaya, 2013). We speculated that subjects' ratings of the amount of category-typical mechanical- or human-resemblance would exhibit a similar delay in response time for stimuli near a potential categorical boundary. Therefore, we assessed the hypothesis that category confusion causes the UV by testing the following predictions:

1. The time required to rate mechanical- or human-resemblance of a face should be greatest for faces closest to the maximal UV effect on likability.
2. A face's position on the MH spectrum should influence its likability indirectly through category confusion. Thus, rating time should statistically mediate the nonlinear relationship between MH score and likability.

2.5.1. Methods

We conducted exploratory post hoc analyses to assess both of these predictions. Using data from Experiment 1A, we used inverse-variance weighted polynomial regression to model median rating time by MH score, using *F*-tests to select the best-fitting model from linear, second-degree, and third-degree models.

We used structural equation modeling to assess whether rating time mediated the relationship between MH score and likability. We modeled likability and time as their best-fitting polynomial functions of MH scores. Based on visual inspection, we modeled the relationship between rating time and likability as linear. Inference for the indirect (mediation) effect of time was based on bias-adjusted bootstrapping.

2.5.2. Results

The second-degree regression modeling time by MH score was the best-fitting (second-degree versus linear: $F(1) = 12.64$, $p < 0.001$; versus third-degree: $F(1) = 2.32$, $p = 0.13$) and was used in subsequent mediation analysis. As predicted by the category confusion hypothesis, the fitted curve ($R^2_{adj} = 0.48$; Fig. 3C, solid line) suggests that subjects rated very mechanical faces most quickly, that rating times increased as faces became more human-like, and that rating times again declined somewhat as faces became very human-like. This is consistent with previous studies that also detected a category boundary in nonhuman–human morphed face series (at roughly 70% human) using a different method, namely perceptual discrimination (Cheetham, Suter, & Jancke, 2014; Looser & Wheatley, 2010). Strikingly, the maximum rating time we observed occurred at almost exactly the same position on the MH spectrum as the point of minimum likability that we estimated in Experiment 1B (MH scores 39 and 36, respectively). However, mediation by time accounted for only a non-significant 3% of the relationship between MH score and likability ($b = -0.02$; $p > 0.25$; 95% CI: -0.09, 0.05). Thus, while peak rating times coincide closely with the location of the UV on the MH spectrum (prediction #1), our mediation analysis (prediction #2) did not indicate a strong contribution of boundary confusion to the UV effect.

3. Experiment 2: a controlled series of composed face images

3.1. Methods

Experiments 2A and 2B shared the following stimuli and subject recruitment procedures. We created stimuli using a complementary approach to that used in Experiment 1. Rather than using a large collection of photographs of actual robots, we precisely controlled a number of extraneous parameters by digitally composing a series of robot faces ranging from very mechanical to very human-like. A series of robot faces varying along the spectrum of robot to human would be most easily created by digitally morphing in stages between a purely robotic and a purely human face using commercial software to cross-dissolve and warp the faces (Katsyri et al., 2015). However, our goal was to study human

[9] The apparent discrepancy between (1) the nonsignificant dichotomized interaction test and (2) the attenuation in the Uncanny Valley effect observed in the low-emotion subset of faces may reflect the fact that there were relatively few low-emotion robots at the most human-like end of the spectrum. The exclusion of these robots in the low-emotion subset could dampen the curvature of the Uncanny Valley at the most human-like end of the spectrum by removing the goodness-of-fit incentive for a nonlinear fit in this region without producing a statistical interaction. The likability outcome did not show this attenuation among low-emotion faces, likely because its more pronounced "valley" justified a higher-degree polynomial fit regardless of a loss of statistical efficiency in the most human-like region.

responses to realistic robot faces that can actually be built and that the subjects believe to exist. Therefore, each face was individually composed (Adobe Photoshop CS) using parts derived from images of actual built robots (as well as one doll and one human) and morphometry was standardized to the human face. We adopted this approach not only to preserve the "buildability" of the faces, but also to help ensure that each face exhibited a set of features considered by working designers to be congruous and appropriate to that type of robot.

We controlled a wide variety of characteristics, including:

1. Framing, angle, and lighting.
2. Aspect ratio of face (and contours, when skin was present).
3. Position, angle, and size of mouth, nose, eyes, eyebrows, and ears (when present).
4. Iris size and color.

Because parameters for realism in a human face are more constrained than those in robot faces, we used the human face as the standard to which the other faces were adjusted. The faces were presented in color on a white background. The resultant six-face series is shown in Fig. 4A. As a manipulation check, we recruited a separate sample of subjects to rank the stimuli in order of perceived human-resemblance; these subjects ($n = 12$) all selected the intended robot-to-human ordering.

We used a printed questionnaire to measure the subjects' emotional responses to each robot face. The questionnaire instructed subjects to view the six face images and rate how much they thought they would like to interact with each. Specifically, subjects were instructed 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." Subjects rated the likability of each robot on a VAS similar to that used in Experiment 1B. If subjects were to have noticed that the faces formed an orderly series progressing from mechanical to human, they might have tended to rank their responses to conform to that pattern. Therefore, we arranged the faces in a randomized order that was the same for all subjects. In addition, to further disguise the close relationship among the faces, "decoy" robot faces (responses to which were not analyzed) were interspersed with the actual test faces. Because the order of presentation (even when randomized) might have influenced subjects' expectations of how widely the faces would range in likability, subjects were instructed to quickly browse through the entire series of faces before beginning the task and were permitted to revise their responses after rating all the faces.

3.2. Experiment 2A: likability

3.2.1. Methods

To investigate the presence of an UV effect, we obtained estimates of the relative likability of each face by fitting a linear mixed-effects model with random intercepts by subject. The primary coefficients of interest were the main effects for each face,

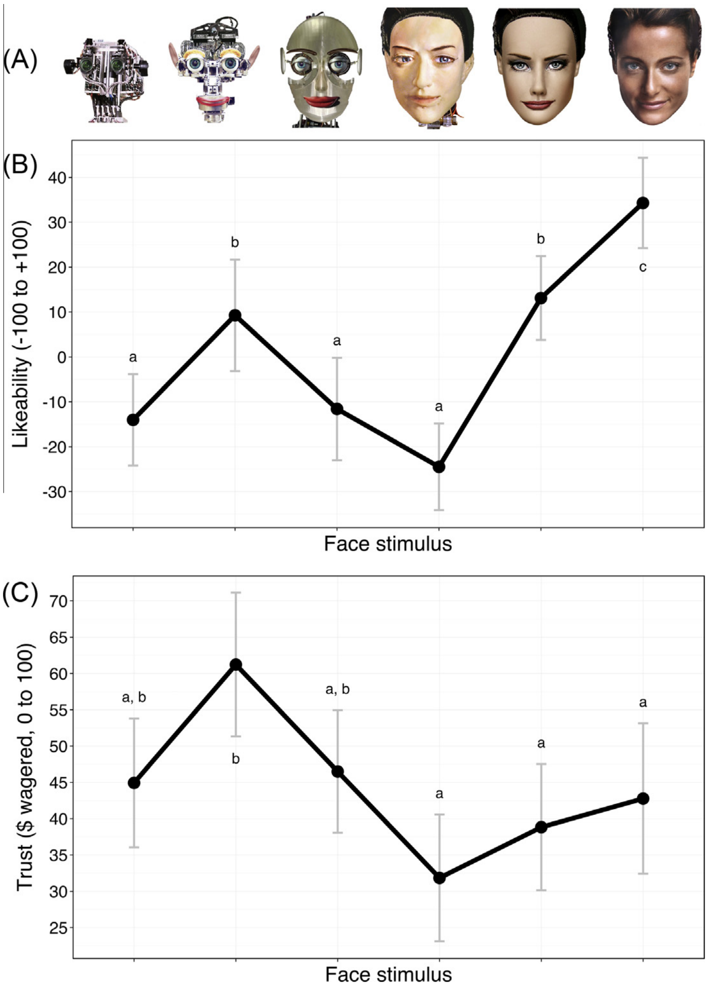

Fig. 4. Uncanny Valley in a controlled face series (panel A) for likability (panel B) and trust-motivated wagering (panel C). Error bars represent 95% confidence intervals. Faces sharing the same letter annotation did not differ significantly from each other on the outcome (based on Tukey-adjusted *t*-tests of least-squares means).

treated as a categorical variable. We used a χ^2 test of nested models to assess for overall differences in likability across the faces. To assess for the specific differences in likability across faces predicted by the UV model, we used Tukey-adjusted, pairwise *t*-tests of least-squares means estimated via the mixed model fit.

3.2.2. Results

Subjects ($n = 52$ after the exclusion of 1 subject who misunderstood how to use the VAS) were 40% male with median age 44 years. There were significant differences in likability across the face stimuli (Fig. 4B; global test of nested models: $\chi^2_5 = 80.63$, $p < 0.001$). It is important to note that intervals between faces on the *X*-axis cannot be assumed to be equal. The fully human Face #6 was rated the most likable, with a mean rating far above the neutral point of the scale (mean VAS rating = +34; 95% CI: 24, 44); Face #4 was rated the least likable, with a mean rating far below the neutral point of the scale (mean VAS rating = −25; 95% CI: −34, −15). Faces #1, #3, and #4 were considered significantly less likable than Faces #2, #5, and #6; the fully human Face #6 was considered significantly more likable than all the others (pairwise comparisons, $p < 0.05$). There were not significant differences in likability across Faces #1, #3, and #4, or between Faces #2 and #5.

3.3. Experiment 2B: trust-motivated behavior

3.3.1. Methods

Subjects completed an investment-game implicit measure of trust as in Experiment 1C; stimuli were identical to those used in Experiment 2A. As in Experiment 2A, we used a linear mixed-effects model to assess for UV effects on trust. Secondarily, we investigated the effect of a subtle manipulation of facial expression, namely a change in eyebrow angle, on inferred trustworthiness (Appendix B).

3.3.2. Results

Subjects ($n = 92$) were 78% female with median age 32 years. Each of the face stimuli was viewed on average by 46 subjects.[10] The median response time was 7.5 s. Fig. 4C shows the results of the investment game. Subjects wagered a mean of $44 of the $100 endowment (95% CI: $41, $48). The mixed-effects model indicated significant differences across faces in trust elicited (global test of nested models: $\chi^2_5 = 28.31$, $p < 0.001$). Face #2 elicited significantly more trust than Faces #4, #5, and #6 ($p < 0.05$), while all other pairwise comparisons were non-significant.

3.4. Replication of experiment 2 with alternative stimuli

3.4.1. Methods

We replicated Experiments 2A–2B using a second novel set of composed stimuli. To provide assurance that these new stimuli were selected *a priori* and were not edited to force a replication of the initially reported effect, stimuli and procedures for the replication study were pre-registered (https://osf.io/3rjnk/). All methods, including analysis, were as described for Experiment 2A–2B, except as follows. The replication took place online using recruitment procedures as in Experiment 1. We used empirical ratings of human- and mechanical-resemblance estimated in Experiment 1 to guide the construction of a new set of six stimuli that appropriately spanned the MH spectrum. Specifically, we informed the appearance of the new stimuli (Fig. 5A) by the faces in Experiment 1A that most closely achieved average MH scores at the 0% (fully mechanical), 20%, 40%, 60%, and 80% points of the scale. We modified an image of an actual human to serve as the new Face #6. As in Experiment 2, face stimuli underwent extensive digital editing in order to standardize morphometry to the human face. A manipulation check ensured that the stimulus set had achieved the intending ordering from mechanical to human-like. (However, it is important to note that, for example, Face #4 in the new stimulus set cannot be assumed to correspond to exactly the same position on the *X*-axis as Face #4 in the original stimulus set.)

3.4.2. Results for likability

105 subjects (no exclusions) completed Replication Experiment 2A. There were significant differences in likability across the face stimuli (Fig. 5B; global test of nested models: $\chi^2_5 = 214.80$, $p < 0.001$). The fully human Face #6 was rated the most likable (mean VAS rating = +43; 95% CI: 32, 54). Face #1 was rated the least likable, (mean VAS rating = −40; 95% CI: −48, −31), followed by Face #3 (mean VAS rating = −33; 95% CI: −41, −26). All Tukey-adjusted pairwise comparisons were significant except for those between Faces #1 and #3 and between Faces #3 and #4.

3.4.3. Results for trust-motivated behavior

98 subjects (after the exclusion of 2 subjects who reported substantial technical or comprehension problems) completed Replication Experiment 2B. There were not significant differences in trust across the face stimuli (Fig. 5C; global test of nested models: $\chi^2_5 = 8.76$, $p = 0.12$), and pairwise comparisons were all nonsignificant.

4. Discussion

Humans often have unexpectedly uncomfortable reactions to android robots that were designed to have pleasant social interactions with humans. The Uncanny Valley theory is a commonly cited, but controversial, explanation for human discomfort with imperfect human likenesses, yet it has been tested empirically by only a few small studies that yielded conflicting results. We used innovative methods to address some of the most difficult issues in measuring human social and affective reactions to social android robot faces.

To minimize the biases that potentially plague the selection and creation of a small set of robot faces to represent the mechanical-human spectrum, our first major innovation was the use of two complementary and novel methods of stimulus creation. In Experiment 1, we used an objective selection process to obtain a large sample of images from the actual wild-type population of robots that have been built under real-world constraints of design and construction for the purpose of social interaction. By empirically estimating the perceived human–mechanical resemblance of each face, we precisely located their positions on a fine-grained spectrum from mechanical to human. Furthermore, separately measuring mechanical- and human-resemblance of the faces enabled us to validate the largely unquestioned assumption that the Uncanny Valley can be reasonably represented on a single, unidimensional *X*-axis of mechano-humanness. Additionally, we addressed limitations of relying solely on self-report as a measure of social reactions to faces. To assess how deeply those explicit reactions penetrate to affect actual social behavior, we quantified trust-motivated behavior by developing a novel adaptation of the classic game-theory paradigm of a one-shot wagering game.

In Experiment 1, as an inherent feature of the stimulus selection process, the robot faces showed natural variability in factors other than human–mechanical resemblance that might influence likability and trust, such as proportions (Stirrat & Perrett, 2010), the presence or absence of various facial features, facial expression (Eckel &

[10] As detailed in the Appendix, due to a secondary experiment nested within Experiment 2B, subjects did not each judge all six faces.

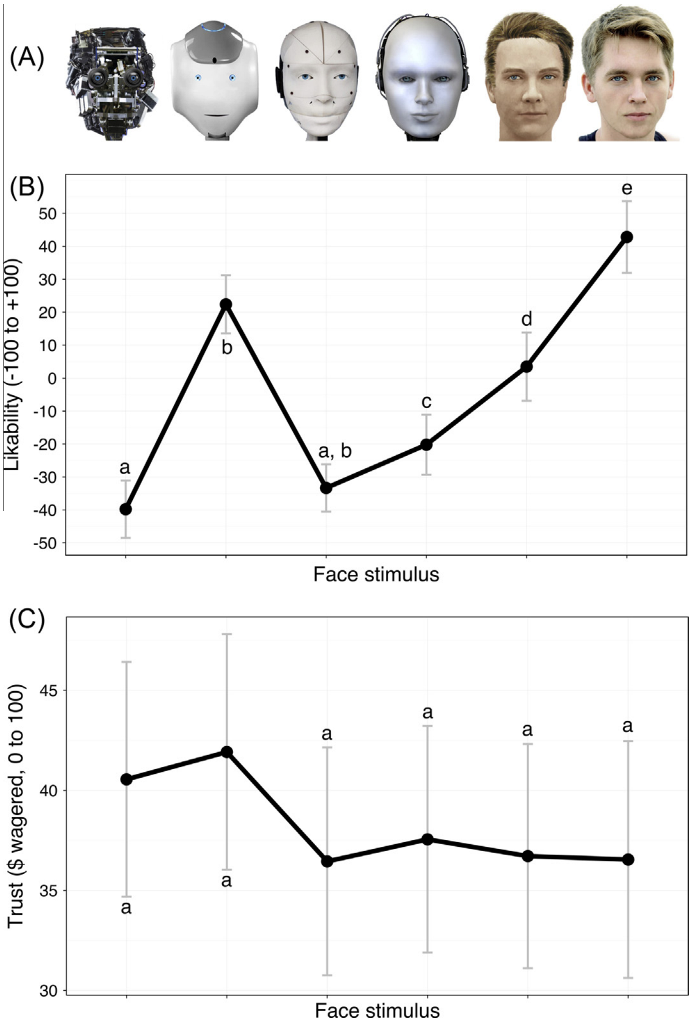

Fig. 5. Uncanny Valley in a replication stimulus set (panel A) for likability (panel B) and trust-motivated wagering (panel C). Error bars represent 95% confidence intervals. Faces sharing the same letter annotation did not differ significantly on the outcome (based on Tukey-adjusted *t*-tests of least-squares means).

Wilson, 1998), displayed emotion (Scharlemann et al., 2001), positioning (Mara & Appel, 2015), background setting, sex (Bohnet & Zeckhauser, 2004; Buchan, Croson, & Solnik, 2008), age (Holm & Nystedt, 2005), race (Baxter, 1973; Glaeser, Laibson, Scheinkman, & Soutter, 2000), ethnicity (Fershtman & Gneezy, 2001), physical attractiveness (Wilson & Eckel, 2006), and resemblance to the viewer (DeBruine, 2002). Additionally, despite our use of objective inclusion criteria, the corpus of robot faces images available through an Internet search may be a biased representation of the total possible range of robots; factors such as designer intentions and intended audiences may confound the relationship between mechano-humanness and elicited social responses.

Therefore, Experiment 2 took a complementary approach in which we digitally created a novel set of robot faces with tightly controlled morphometry, expression, and presentation. These stimuli were based on real androids (and one human) in order to respect real-world constraints of robot build and design (in contrast to the unnaturally blended images or amalgams used in previous studies).

4.1. An Uncanny Valley for explicitly-rated likability

We found that all key characteristics of the Uncanny Valley are robustly apparent in both the wild-type sample and digitally composed robot face stimuli. To a point, likability increased with increasing human-resemblance beyond the nearly neutral reactions elicited by the most mechanical robots. But as faces became more human than mechanical, they began to be perceived as frankly unlikable. Finally, as faces became nearly human, likability sharply rebounded to a final positive ending point. This steep final increase is an important feature in Mori's conceptualization (Mori, 1970) and suggests that although the most human-like robots may be more likable than reliably perfect human likenesses, they may occupy a precarious position at which small faults in their humanness might send the social interaction tumbling. The results of Experiment 2 confirm that for a particular morphometric configuration of robot face, an Uncanny Valley effect can indeed be a potent factor. Results were strikingly similar between the wild-type and composed stimulus sets, both in qualitative visual patterns and in the quantitative likability values defining the inflection points – that is, the actual depth of the Uncanny Valley's nadir and height of its apices. In contrast, a previous study (whose sample of wild-type robot whole-body images included non-human characters and contained few highly human-like robots) that were rated on a 5-point Likert scale did not find evidence of an Uncanny Valley (Rosenthal-von der Pütten & Krämer, 2014).

Of course, the Uncanny Valley is not the only important factor in determining a robot's perceived likability: in Experiment 1, there were individual wild-type robots throughout the mechano-humanness spectrum that were much more likable or unlikable than predicted by the Uncanny Valley effect alone. Perceived emotion, for example, may play a central role; predictably, robots showing more positive emotion were perceived as more likable. However, the Uncanny Valley effect persisted among faces perceived as displaying almost no emotion.

Previous theories have cited category confusion occurring at the mechanical-human boundary as a possible mechanism for Uncanny Valley effects (Cheetham et al., 2014; Katsyri et al., 2015). Indeed, our preliminary analysis indicated that the position on the mechano-humanness spectrum associated with the longest rating times in Experiment 1A (suggestive of confusion or deliberation) coincided almost exactly with the point of minimum likability in Experiment 1B. However, rating time did not statistically mediate the relationship between mechano-humanness and likability, highlighting the need for more research into mechanisms underlying the Uncanny Valley. Such research might inform practical design choices to circumvent Uncanny Valley effects or might alternatively indicate that the Uncanny Valley is an inherent and insurmountable feature of human category perception.

4.2. An Uncanny Valley for trust-motivated behavior

Similarly to explicit ratings of likability, trust-motivated behavior with real financial consequences demonstrated all the key features of an Uncanny Valley in both the wild-type sample and the controlled face series. However, in the post hoc replication using a second controlled face series, trust-motivated behavior did not exhibit pronounced Uncanny Valley characteristics. Of note, the original controlled series of faces (Experiment 2B) represented a female individual, while the replication stimulus series (Fig. 5A) was male, suggesting that trust may be particularly sensitive to differences in the characteristics of the individual represented by the series. Specifically, the apparently attenuated Uncanny Valley for trust in the male replication stimuli qualitatively appears to largely reflect a failure of the most human-like faces to "recover" from an initial small decrease in trust relative to the most mechanical robot faces in the series. Given that the sex of a robot may be more obvious in human-like versus mechanical faces, this attenuation may reflect a known tendency for male humans to be perceived as less trustworthy than female humans (Bohnet & Zeckhauser, 2004; Buchan et al., 2008). We speculate that, compared to likability, Uncanny Valley effects on trust may interact more strongly with robots' individual characteristics including, for example, sex, facial morphometry, and perceived emotion. Identifying such moderators in future research would clarify both cognitive mechanisms and practical impacts of Uncanny Valley effects on trust.

Collectively, these results suggest that the Uncanny Valley can be more than a superficial result of asking subjects to consciously judge the robots' likability – rather, at least in some types of robots, it has more profound effects on the emotional-cognitive motivations of strategic social behavior, affecting one of the most fundamental social judgments: that of trustworthiness (Cosmides & Tooby, 1992; de Melo, Carnevale, & Gratch, 2013).

4.3. What is meant by "trusting" a robot?

We have used the term "trust" to describe subjects' willingness to take a real financial risk on the possibility that a robot will act prosocially in the subject's interest. Indeed, the "investment game" model (Berg et al., 1995) from which our game derives is generally accepted as the standard experimental instrument for measurement of trust (Camerer, 2003). One important factor in such differential judgments of the trustworthiness of individuals is the concept of "encapsulated interest" (Hardin, 2002); that is, we trust others whom we believe have interests that encapsulate our own. The idea that our subjects may have attempted to assess the robots' "interests" raises the intriguing possibility that our measurements touch upon the subjects' application of a theory of mind (Premack & Woodruff, 1978) to their robot social partners, to some degree attributing an intentional stance to the robots that implies thoughts, beliefs, and desires (Dennett, 1989). Alternatively, it is possible that the subjects' decisions regarding encapsulated interest were not, in fact, directed toward the robot, but rather that they treated the robot face as an inanimate agent of the robot maker. In Searle's terms (Searle, 1983), the subjects metaphorically rationalized their interaction with the robots by attributing an "as-if intentionality" to them, but this is distinct from the true "intrinsic intentionality" of the human designer. Experiments employing game-theory methods can be designed to address these issues in the future; an enriched theoretical understanding of trust toward robots could inform robot design choices to overcome Uncanny Valley effects.

4.4. Conclusions

Our investigations indicate that the Uncanny Valley is a real influence on humans' perceptions of robots as social partners, robustly influencing not only humans' conscious assessments of their own reactions, but also able to penetrate more deeply to modify their actual trust-related social behavior with robot counterparts. In addition, for robots throughout the mechano-humanness spectrum, humans appear to infer trustworthiness from affective cues (subtle facial expressions) known to govern human–human social judgments. These observations help locate the study of human-android robot interaction squarely in the sphere of human social psychology rather than solely in the traditional disciplines of human factors or human–machine interaction (Hoff & Bashir, 2015). Our innovative methods of assessing human social perceptions of android robots in relation to their degree of mechano-humanness provide tools for further studies into social psychological and affective factors that could inform the design of socially competent robots.

Data transparency

All raw data, R code, and questionnaire materials (including files that can be directly imported into Qualtrics) are publicly available at https://osf.io/3rjnk/.

Author contributions

Both authors conceptualized the project, contributed to study design, collected data, and wrote the manuscript. MM planned and performed all statistical analyses. Both authors approved the final version of the manuscript.

Acknowledgments

This research had no external funding. We are deeply grateful to Dr. Lynn Baroff for many insightful discussions and encouragement.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.09.008.

References

Baxter, G. W. (1973). Prejudiced liberals? Race and information effects in a two-person game. *Journal of Conflict Resolution, 17*, 131–161.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122–142.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Economic Behavior and Organization, 55*, 467–484.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems, 42*(3), 167–175.

Buchan, N. R., Croson, R. T. A., & Solnik, S. (2008). Trust and gender: An examination of behavior and beliefs in the investment game. *Journal of Economic Behavior and Organization, 68*, 466–476.

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

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

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.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkhow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind and the generation of culture* (pp. 163–228). Oxford: Oxford University Press.

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

de Melo, C., Carnevale, P., & Gratch, J. (2013). People's biased decisions to trust and cooperate with agents that express emotions. In *Proceedings from international conference on autonomous agents and multiagent systems*.

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proceedings of the Royal Society of London B: Biological Sciences, 269*(1498), 1307–1312.

Dennett, D. C. (1989). *The intentional stance*. Cambridge: MIT Press.

Eckel, C. C., & Wilson, R. K. (1998). Reciprocal fairness and social signalling: Experiments with limited reputations. In *Proceedings from American Economic Association Meeting, New York, NY*.

Fershtman, C., & Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics, 116*, 351–377.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robots and Autonomous Systems, 42*, 143–166.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *Quarterly Journal of Economics, 115*, 811–846.

Hanson, D. (2006). Exploring the aesthetic range for humanoid robots. In *Proceedings from ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science*.

Hardin, R. (2002). *Trust and trustworthiness*. New York, NY: Russell Sage Foundation.

Hoff, K. A., & Bashir, M. (2015). Trust in automation integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(3), 407–434.

Holm, H., & Nystedt, P. (2005). Intra-generational trust—A semi-experimental study of trust among different generations. *Journal of Economic Behavior and Organization, 58*, 403–419.

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

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., 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.

Mailath, G. J., & Samuelson, L. (2006). *Repeated games and reputations*. Oxford: Oxford University Press.

Mara, M., & Appel, M. (2015). Effects of lateral head tilt on user perceptions of humanoid and android robots. *Computers in Human Behavior, 44*, 326–334.

Mori, M. (1970). The Uncanny Valley. *Energy, 7*(4), 33–35 [Japanese].

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics, 15*(2), 285–290.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526.

Reips, U. D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS generator. *Behavior Research Methods, 40*(3), 699–704.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (pp. 243–335). New York: Academic Press.

Rosenthal-von der Pütten, A. M., & Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. *Computers in Human Behavior, 36*, 422–439.

Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology, 22*(5), 617–640.

Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science, 21*(3), 349–354.

Wilson, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly, 59*, 189–202.

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.

Zhao, S. (2006). Humanoid social robots as a medium of communication. *New Media & Society, 8*, 401–419.