Neural correlates of the uncanny valley effect for robots and hyper-realistic masks

Shona Fitzpatrick[1], Ailish K. Byrne[2], Alex Headley[1], Jet G. Sanders[3], Helen Petrie[4], Rob Jenkins[1] & Daniel H. Baker[1,5]

1. Department of Psychology, University of York, York, UK
2. School of Medicine, Keele University, Newcastle-under-Lyme, Staffordshire, UK
3. Department of Psychological and Behavioural Science, London School of Economics, London, UK
4. Department of Computer Science, University of York, York, UK
5. Corresponding author, email: daniel.baker@york.ac.uk

1 Abstract

Viewing artificial objects and images that are designed to appear human can elicit a sense of unease, referred to as the 'uncanny valley' effect. Here we investigate neural correlates of the uncanny valley, using still images of androids (robots designed to look human), and humans wearing hyper-realistic silicone masks, as well as still images of real humans, in two experiments. In both experiments, human-like stimuli were harder to distinguish from real human faces than stimuli that were clearly not designed to mimic humans but contain facial features (mechanical robots and Halloween masks). Stimulus evoked potentials (electromagnetic brain responses) did not show convincing differences between faces and either androids or realistic masks when using traditional univariate statistical tests. However, a more sensitive multivariate analysis identified two regions of above-chance decoding, indicating neural differences in the response between human faces and androids/realistic masks. The first time window was around 100-200ms post stimulus onset, and most likely corresponds to low-level image differences between conditions. The second time window was around 600ms post stimulus onset, and may reflect top-down processing, and may correspond to the subjective sense of unease characteristic of the uncanny valley effect. Objective neural components might be used in future to rapidly train generative artificial intelligence systems to produce more realistic images that are perceived as natural by human observers.

2 Introduction

Many people report an aversion to entities that are superficially human-like, but on closer inspection turn out to be artificial. Examples include humanoid robots (androids), puppets, hyper-realistic masks [1], and computer-generated images or movies. The term 'uncanny valley' [2,English translation in 3] describes the idea that clearly human or clearly artificial entities do not evoke unease, whereas artificial entities that are human-like are disconcerting. Understanding these experiences is increasingly important as artificial entities become more integrated into our everyday lives. However at present relatively little is known about the neural underpinnings of the uncanny valley effect [for a recent review, see 4]. In particular, the root of the uncanny valley effect remains debated: does it arise primarily from bottom-up sensory conflicts, or from higher-level cognitive processes? Resolving this question is critical to understanding its fundamental mechanisms.

Neural responses to faces and bodies in general are well-characterised, and there appear to be specialised brain regions devoted to both [reviewed in 5]. For example, areas of the occipital lobe [6] and fusiform gyrus [7] respond more to faces than non-face stimuli, and sections of extrastriate cortex are responsive to bodies [8]. There are also electromagnetic event-related potential (ERP) signals associated with face and body stimuli, though their precise role is still debated [9,10]. It seems highly likely that 'uncanny' images will activate these same processes, yet it is unclear whether the initial cause of the sense of unease they produce occurs at bottom-up sensory stages [11–13] or is modulated by more top-down cognitive factors [14,15].

One previous study by [16] measured functional magnetic resonance imaging (fMRI) responses to moving stimuli designed to elicit an uncanny valley effect. They found repetition suppression effects[1] in action-specific brain regions responding to movies of androids that had a biological appearance, but mechanical motion. These effects were stronger than for movies of humans or mechanical robots performing the same actions. A more recent electroencephalography (EEG) study [17] identified a difference in the N400 component[2] between dynamic and static conditions using the same stimuli. Although this difference was strongest over frontal electrodes, source reconstruction of the N400 itself suggested a left-lateralised source in the temporo-parietal cortex, consistent with the fMRI results [16]. The authors interpret both of these findings as being due to the discrepancy between the human-like appearance and the clearly non-biological motion of the robot.

Our aim was to further investigate neural correlates of the uncanny valley effect, with the expectation that increased understanding will aid efforts to generate more convincingly human robots and avatars in the future. We achieved this through two EEG experiments, in which we measured neural responses to static images. Although previous studies focus on dynamic stimuli, static images allow for a more precise investigation of the neural mechanisms underlying the uncanny valley effect, particularly by eliminating motion-related confounds. In the first experiment, the stimuli were still images of humans, machine-like robots, and human-like robots (see Figure 1a). In the second experiment we aimed to generalise the finding by using images of people wearing no masks, wearing obvious masks (e.g. Halloween masks), and wearing hyper-realistic silicone masks [18] (see Figure 1b). Rather than focus on specific ERP components, we use a multivariate pattern classification approach (a machine learning technique in which an algorithm is trained to decode the neural responses) to identify time windows in which information in the EEG signal can be used to distinguish between pairs of conditions. Our rationale is that timepoints where signals evoked by human faces can be distinguished from those evoked by human-like robots, or hyper-realistic masks, are candidates for a neural signature of the uncanny valley effect.

3 Materials & Methods

3.1 Participants

A total of 29 participants completed Experiment 1 (12 male, 17 female), and 30 different participants completed Experiment 2 (7 male, 23 female). Participants were young adults with no history of neurological disorder. None of the participants had previously taken part in a study using these stimuli, and all were naïve to the hypotheses and wore their normal optical correction if required. Written informed consent was collected before each experiment began, and all procedures were approved by the Ethics committee of the Department of Psychology at the University of York. Data collection for Experiment 1 ran from 14th July to 15th September 2022, and data collection for Experiment 2 ran from 12th October 2017 to 14th February 2018.

3.2 Apparatus & stimuli

In Experiment 1, the stimulus set consisted of a total of 90 images, evenly split between three categories: real faces, human-like robots, and mechanical robots. Images all showed the head and shoulders of the subject, had white backgrounds, and were sourced from the Internet. In Experiment 2, the stimulus set (first described by [19], but here including additional images) consisted of a total of 296 images, comprising real faces (148 images), people wearing silicone masks (74 images), and people wearing obvious masks of the sort typically worn for carnivals and Halloween celebrations (74 images). The backgrounds of these images were more heterogeneous, and showed the natural surroundings of the subject. While the image backgrounds

[1] Repetition suppression is a phenomenon in which the neural response to repeated presentations of identical or similar stimuli is reduced relative to the response on the first presentation.

[2] The N400 is an electromagnetic brain potential obtained 400ms after stimulus onset, typically over centro-parietal electrodes. It has been proposed to reflect the extent to which the stimulus presented was surprising or unexpected.

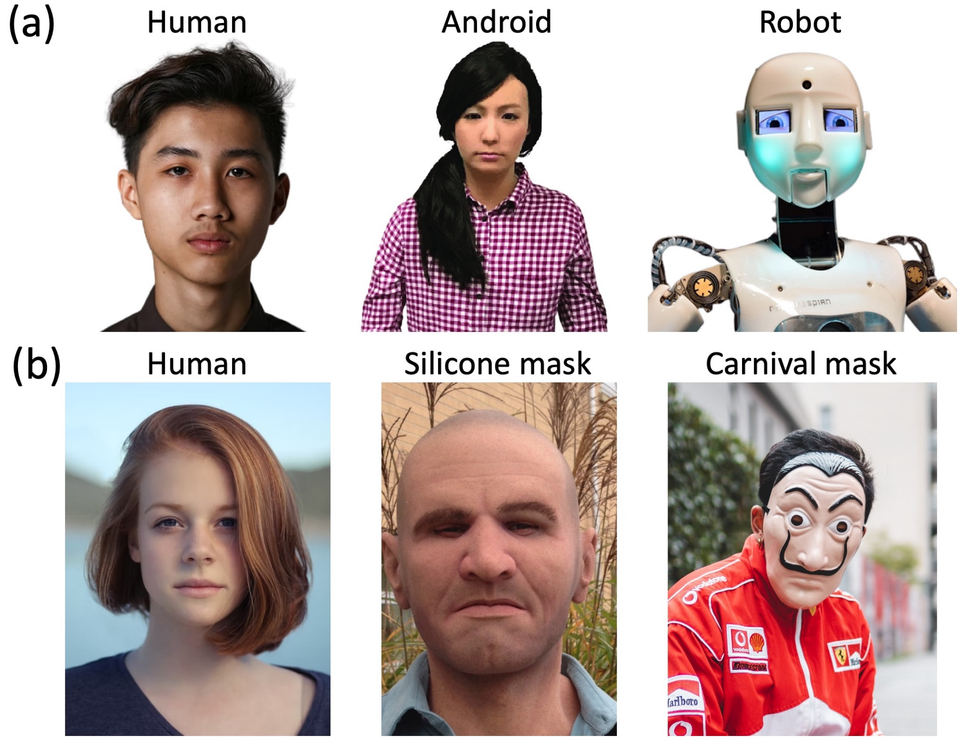

Figure 1: Illustrative stimuli from the same categories as used in Experiments 1 and 2. Row (a) shows a human face, an android and a robot, all against white backgrounds. Row (b) shows a human face, a hyper-realistic silicone mask, and a Halloween mask, all against natural backgrounds. Images shown here were taken from a variety of sources that permit reuse in academic contexts and in most cases were not part of the stimulus set from the experiments. The silicone mask image was taken by the authors (subject: RJ, image credit: JGS), and was used in Experiment 2. The individual in this manuscript has given written informed consent (as outlined in PLOS consent form) to publish these case details (i.e. this image).

differed across experiments, we hypothesize that the primary task was not affected, as participants focused on the foreground stimuli. In both experiments, images included examples of both genders, and of varied ethnic backgrounds.

All stimuli were displayed on a ViewPixx display running at 120Hz, controlled by an Apple Macintosh computer. The display was gamma corrected using a photometer to ensure that the luminance output was linear. EEG data were collected using a 64-channel Waveguard cap and an ANT Neuroscan system, sampling at 1kHz. The ground electrode was located at position AFz, and all signals were referenced to the whole head average. Low latency digital triggers were sent between the display and the EEG amplifier using an 8-bit parallel cable.

3.3 Procedure

3.3.1 Experiment 1: Robots

Each participant completed three blocks of the first experiment. Within each block, all 90 stimulus images were presented twice in a random order. Stimuli subtended 11×11 degrees at the viewing distance of 57cm, and were shown against a mid-grey background, with a black central fixation cross displayed throughout. The presentation duration was 500ms, and participants were asked to press a mouse button to indicate if they believed each image was of a human or of a robot/android. After each response there was a random duration blank period with a mean duration of 1000ms and a standard deviation of 200ms. Durations were chosen to provide sufficient information for judgment while avoiding task fatigue. Randomized blank periods were designed to reduce carryover effects and prevent anticipatory biases. Each block lasted around 6 minutes.

After the EEG experiment, participants also completed a series of questionnaires using the Qualtrics platform. These involved rating their perception of a subset of the stimuli (8 from each category), using items from the Godspeed questionnaire [20]. Items were selected that were expected to be most closely aligned to measuring the sensation of uncanniness. Inspection time was unlimited. Participants also provided demographic information (age, gender) and completed the GAToRS [21] and AQ [22] questionnaires, however the results of these additional questionnaires are not presented here.

3.3.2 Experiment 2: Hyper-realistic masks

Participants were shown all 296 images in a random order in each of three blocks. In the first block, stimuli subtended 5.5×7.5 degrees of visual angle when viewed at a distance of 57cm. In the second block, stimuli doubled in size (width and height), and subtended 11×15 degrees at the same viewing distance. In the third block, stimuli doubled in size again, and subtended 22×30 degrees. The rationale for the size manipulation was to investigate whether increasing levels of detail made the silicone masks more identifiable [18,19]. However as that is not the main focus of the current paper, and our preliminary analyses indicated no differences between size conditions, we collapse results across size conditions. Stimuli were presented for 250ms, and participants indicated whether they thought each image contained a real face or a mask, using a two-button trackball. The button assignment (whether the left button indicated a face or a mask, and vice versa) was determined randomly for each participant, but remained constant throughout the whole experiment. Text reminding the participant of the button assignment was present continuously in the lower right corner of the screen, far from the area of the screen where the stimuli were presented. A central fixation cross was also present throughout. After each response there was a random duration blank period with a mean duration of 1000ms and a standard deviation of 200ms. Each block lasted around 8 minutes.

An independent group of 20 participants also completed an online questionnaire in which they rated the images along various dimensions. The participants repeated the real face vs mask judgement from the main experiment, and were additionally asked to rate emotional expressiveness, realism, and uncanniness for each image using a 7-point Likert scale. Inspection time was unlimited for these judgments.

3.4 Data analysis

We analysed response data by calculating d-prime (d') scores for each condition, derived from the hit rate and false alarm rate [23]. For the human conditions, the hit rate was the proportion of human images correctly identified as human, and the false alarm rate was the proportion of robot or mask images that were incorrectly judged as being human. For the robot and mask conditions, the hit rate was the proportion of robot/mask images correctly identified as not being human, and the false alarm rate was the proportion of human images that were incorrectly judged as being non-human (note that this means the false alarm rate was the same for the robot and android conditions, and for the silicone and Halloween mask conditions). We capped infinite d-prime values (which occur e.g. when the hit rate is 1) at an arbitrary ceiling of 5 to prevent outliers from skewing the results, following established conventions in signal detection theory. We log transformed the reaction times (which typically have positive skew) and performed all averaging and statistical analysis on the logarithmic values.

EEG signals were recorded during each block and saved to disc for subsequent offline analysis. We used components of the EEGlab toolbox [24] to import the data into Matlab and collate data across blocks. We then used Brainstorm [25] to filter the data using a bandpass filter (0.5 to 30Hz), epoch by condition, and subtract a pre-trial baseline (the mean voltage for the 200ms before stimulus onset). Five participants were excluded from the EEG analysis of each experiment due to excessive noise. Our attempts to clean up the data from these participants using independent components analysis were unsuccessful. Their behavioural data were unaffected, and are therefore still included in the analysis.

We performed univariate analyses by conducting Bayesian t-tests [26] between ERPs from pairs of conditions at each time point using a JZS prior, using signals pooled across electrodes P6, P8, PO6 and PO8, which are typically associated with visual responses to faces. The resulting Bayes factor score is a summary of the evidence in favour of either the null hypothesis (that the waveforms are equal) or the alternative hypothesis (that they differ). We use the heuristics proposed by Jeffreys [27] that Bayes factors >3 ($log_{10}BF_{10} > 0.5$) constitute some evidence supporting the alternative hypothesis, factors >10 ($log_{10}BF_{10} > 1$) constitute strong evidence, and factors >30 ($log_{10}BF_{10} > 1.5$) constitute very strong evidence.

Multivariate pattern analysis (MVPA) was conducted by training a linear support vector machine algorithm [LibSVM, 28] to discriminate between patterns of activity across electrodes at a specific time point. MVPA is a statistical technique that involves training a machine learning algorithm to identify patterns in data, and then testing its accuracy at classifying unseen data; in EEG analysis above-chance classification is considered evidence of distinct patterns of neural activity between two conditions. Previous work has indicated that EEG data do not typically require more complex nonlinear algorithms [29]. The patterns came from the human face condition and one of the other conditions, for a single participant. Four examples of each pattern were calculated by averaging over random subsets of 20% of the available trials from a given condition, and these were used to train the classifier. The accuracy of the classifier was tested on the remaining trials (that were not used in training) for each condition. This process was repeated 1000 times with different trial permutations to obtain an average accuracy, where chance performance is at 50% correct. The analysis was carried out at all time points, and for each participant separately. We then averaged classifier accuracy across participants, and calculated one sample Bayesian t-tests comparing to chance performance at each time point as described above.

3.5 Data and code availability

Raw data, processed data, and analysis scripts are freely available through the project repository at: https://doi.org/10.17605/OSF.IO/5NZ2H

4 Results

4.1 Experiment 1

We first explored the behavioural results for identification of human versus non-human stimuli. We calculated d-prime scores to compare sensitivity across conditions. Sensitivity was highest for identifying robots ($d' = 4$), but still well above chance for both the human ($d' = 2.72$) and android ($d' = 2.27$) conditions. The Bayes factor score for a one-way ANOVA comparing these three conditions indicated very substantial evidence ($log_{10}BF_{10} = 6.67$) for a difference between conditions, as illustrated in Figure 2a. Pairwise Bayesian t-tests between conditions indicate very convincing differences in sensitivity between robots and androids ($log_{10}BF_{10} = 7.15$) and robots and humans ($log_{10}BF_{10} = 6.27$). The difference between androids and humans ($log_{10}BF_{10} = 5.58$) was also very substantial. The higher d' values for robots (Figure 2a) could indicate that these stimuli are more visually salient.

Reaction times also differed between conditions, though the effects were rather smaller. Reactions were fastest for identifying robots (RT = 688ms), compared with humans (RT = 803ms) and androids (RT = 809ms). The Bayes factor score for a one-way ANOVA comparing these three conditions indicated strong evidence ($log_{10}BF_{10} = 1.44$) for a difference between conditions, as illustrated in Figure 2b. Pairwise Bayesian t-tests between conditions indicate very convincing differences in sensitivity between robots and androids ($log_{10}BF_{10} = 5.01$) and robots and humans ($log_{10}BF_{10} = 4.61$), whereas the reaction time was equivalent between androids and humans ($log_{10}BF_{10}$ = -0.68).

EEG activity showed a clear visually evoked potential over posterior electrode sites (see Figure 2c), with typical components found in response to visual stimuli pooled across all conditions (the P100, N170 and P200 are indicated in the figure). Pairwise comparisons of conditions are shown in Figure 3a,b. In general there is a tendency for the ERP response to human faces to diverge slightly from the other two conditions [30], however the evidence for this divergence was not compelling. Bayes factors exceeded 3 for only a small number of time points around 300-400ms in the comparison between human and robot images (see yellow bars at y = -8 in Figure 3a), but these differences were small considering the variance in the data.

We also conducted multivariate pattern analysis independently at each time-point for the same two comparisons. The evoked responses for human and robot images caused sufficiently distinct patterns of voltages across the scalp that the pattern classifier could distinguish between them from around 100 ms following stimulus onset, with accuracy up to 72% correct (see Figure 3c). Bayes factors exceeded 30 for much of the time window between 100 and 800 ms, indicating that the decoding was meaningfully above chance performance (50% correct). It was also possible to classify between human and android images (see Figure 3d), however performance was much poorer, with a maximum of 62% correct. Classification accuracy had an initial peak around 100ms that provided compelling evidence for above chance classification (BF>30), and a later region of above-chance classification between 500 and 700ms. The early time window likely reflects rapid processing of low-level visual features, such as edges or color contrasts, consistent with P100 and N170 components, whereas the later time window may involve higher-level cognitive processes, such as evaluating emotional content or judging authenticity. In the Discussion we speculate that these two time periods might correspond to distinct types of signal associated with the uncanny valley. More generally, the high classification accuracy for human vs. robot stimuli may be attributed to the salient mechanical elements of robots, whereas the lower accuracy for humanoids reflects their ambiguous human-like appearance, leading to confusion.

Finally, we analysed the rating data from a set of 10 questionnaire items for 8 stimuli from each category. The results are summarised in Figure 4, and in general show differences between stimulus categories along most dimensions. Of particular note, the U-shaped function predicted by the uncanny valley effect was apparent for ratings along the Dislike-Like (Figure 4d), Unfriendly-Friendly (Figure 4e), and Anxious-Relaxed (Figure 4i) dimensions. These are all dimensions with emotional valence, indicating support for the 'uncanniness' of our android stimuli. However we note that the android and robot categories were typically rated as being more similar to each other than to the real human faces. Following Experiment 1, we sought to generalise our results to a different stimulus set, and next report the results of Experiment 2 which used hyper-realistic silicone masks.

4.2 Experiment 2

The results of Experiment 2 were similar to those of Experiment 1, despite using a quite different stimulus set involving images of humans wearing masks, rather than robots. Sensitivity was highest for identifying Halloween masks ($d' = 4.01$), but still well above chance for both the human ($d' = 2.52$) and silicone mask ($d' = 1.29$) conditions. The Bayes factor score for a one-way ANOVA comparing these three conditions indicated very substantial evidence ($log_{10}BF_{10} = 13.34$) for a difference between conditions, as illustrated in Figure 5a. Pairwise Bayesian t-tests between conditions indicate very convincing differences in sensitivity be-

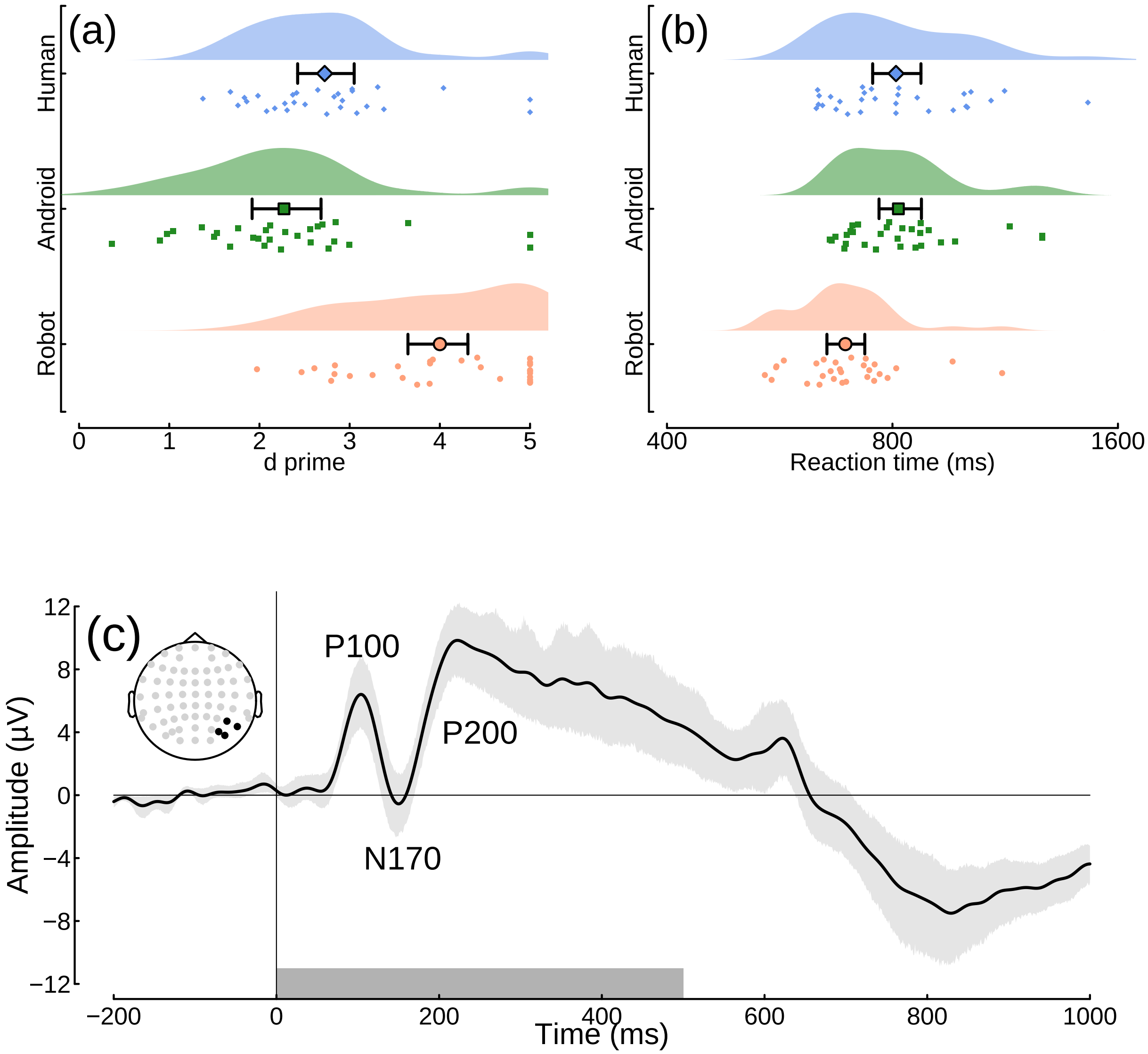

Figure 2: Summary of response data and grand mean ERP for Experiment 1. Panel (a) shows d-prime scores for identifying images of human (blue diamonds), android (green squares) and robot (red circles) faces. Small points show individual participants, and the larger symbols with error bars indicate the group mean and bootstrapped 95% confidence intervals. Panel (b) plots reaction times in the same format (note the logarithmic x-axis). Panel (c) shows the grand mean ERP across all participants and conditions, pooled across electrodes P6, P8, PO6 and PO8 (see inset). The P100 is a positive evoked potential occuring around 100ms after onset of a visual stimulus, associated with the initial (low level) visual response; the N170 is a negative potential at 170ms that is often associated with faces; the P200 is a further positive potential linked to attention and stimulus discrimination. The shaded region around the curve illustrates the 95% confidence interval, and the grey rectangle at the foot indicates the stimulus duration.

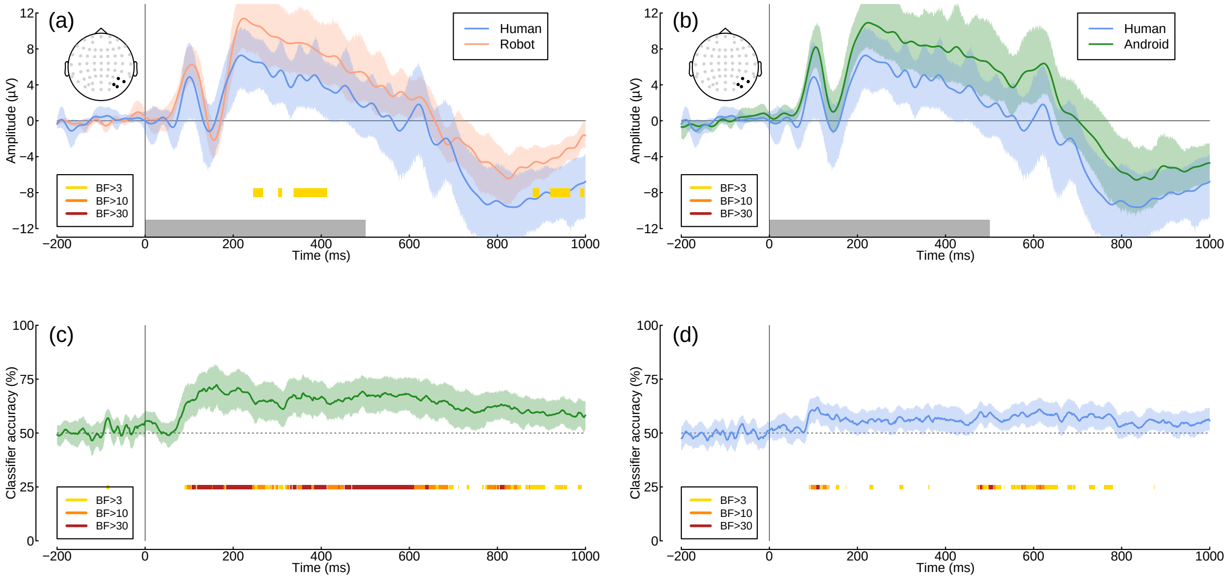

Figure 3: Univariate and multivariate comparisons across image type. Panel (a) shows the ERPs comparing human (blue) and robot (red) face images, and panel (b) compares human (blue) and android (green) faces. Panels (c) and (d) show multivariate pattern classification accuracy for the same comparisons. Points at y = -8 and y = 25 indicate Bayes factor scores for comparisons between ERPs (a,b) and comparing classification accuracy to chance (50% correct; c,d).

tween Halloween and silicone masks ($log_{10}BF_{10} = 9.71$) between Halloween masks and humans ($log_{10}BF_{10} = 11.59$), and between silicone masks and humans ($log_{10}BF_{10} = 5.19$). Unlike in Experiment 1, there were no convincing reaction time differences between conditions ($log_{10}BF_{10} =$ -0.8), as illustrated in Figure 5b.

The grand average ERP waveform for Experiment 2 (see Figure 5c) had similar initial components as for Experiment 1. The latter portion of the waveforms differed somewhat, most likely owing to the difference in presentation duration across experiments (250ms versus 500ms). There was a substantial univariate difference in ERP response between human and Halloween mask conditions extending from around 170 to 230ms following stimulus onset (see Figure 6a), with Bayes factors exceeding 30. Univariate differences between the human and silicone mask conditions were not compelling (see Figure 6b).

Multivariate pattern analysis revealed extremely high classification accuracy (up to 97% correct) comparing human faces with Halloween masks. This was convincingly above chance, with a Bayes factor score exceeding 30 from around 100ms following stimulus onset, and extending across the full time window (see Figure 6c). Classification was also convincingly above chance when comparing human faces with silicone masks (Figure 6d). This timecourse had an initial peak of high accuracy (up to 76% correct) between 100 and 200ms after stimulus onset, followed by a second peak around 600ms. This replicates the finding from Experiment 1 that uncanny valley responses might involve two distinct components at different moments in time.

We subsequently obtained ratings from an independent sample of 20 participants using the same stimuli as in the EEG experiment. This time, we asked for explicit ratings of emotional expressiveness, realism, and uncanniness, as well as repeating the binary real face vs mask rating. Real faces were rated highest for emotional expressiveness (M=3.9) and realism (M=5.9), and lowest for uncanniness (M=2.7) (Figure 7a-c). The realistic silicone masks were rated highest for uncanniness (M=4.5), however this was not dramatically higher than for the Halloween masks (M=4.3). Arguably making judgements of uncanniness is less appropriate for masks that are not intended to be realistic, though our data do qualitatively conform to the U-shaped function expected by the uncanny valley hypothesis. The pattern of d-prime scores (Figure 7d) was similar to those obtained in the main experiment (Figure 5b), with generally higher scores attributable

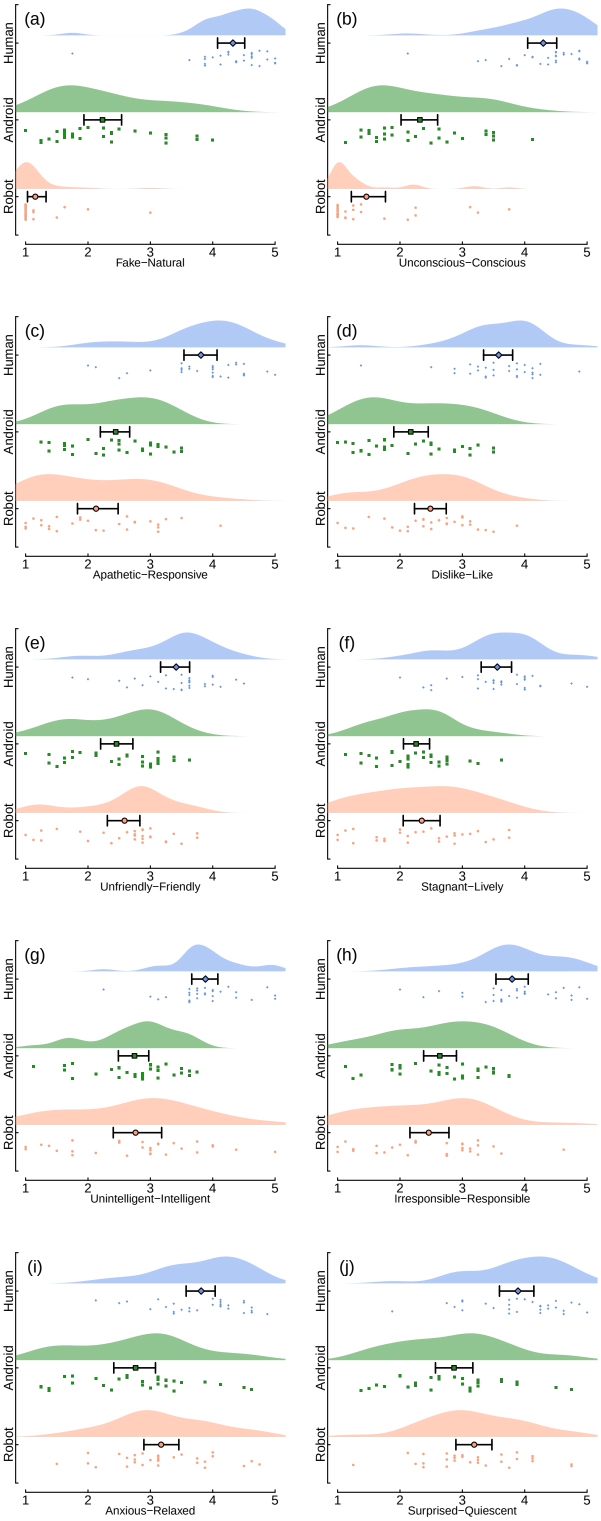

Figure 4: Ratings of stimuli using items from the Godspeed questionnaire. Each rating was on a Likert scale from 1-5, and was the average of ratings from 8 stimulus examples. Dots show individual participant scores, with the larger symbols indicating the mean and 95% bootstrapped confidence intervals.

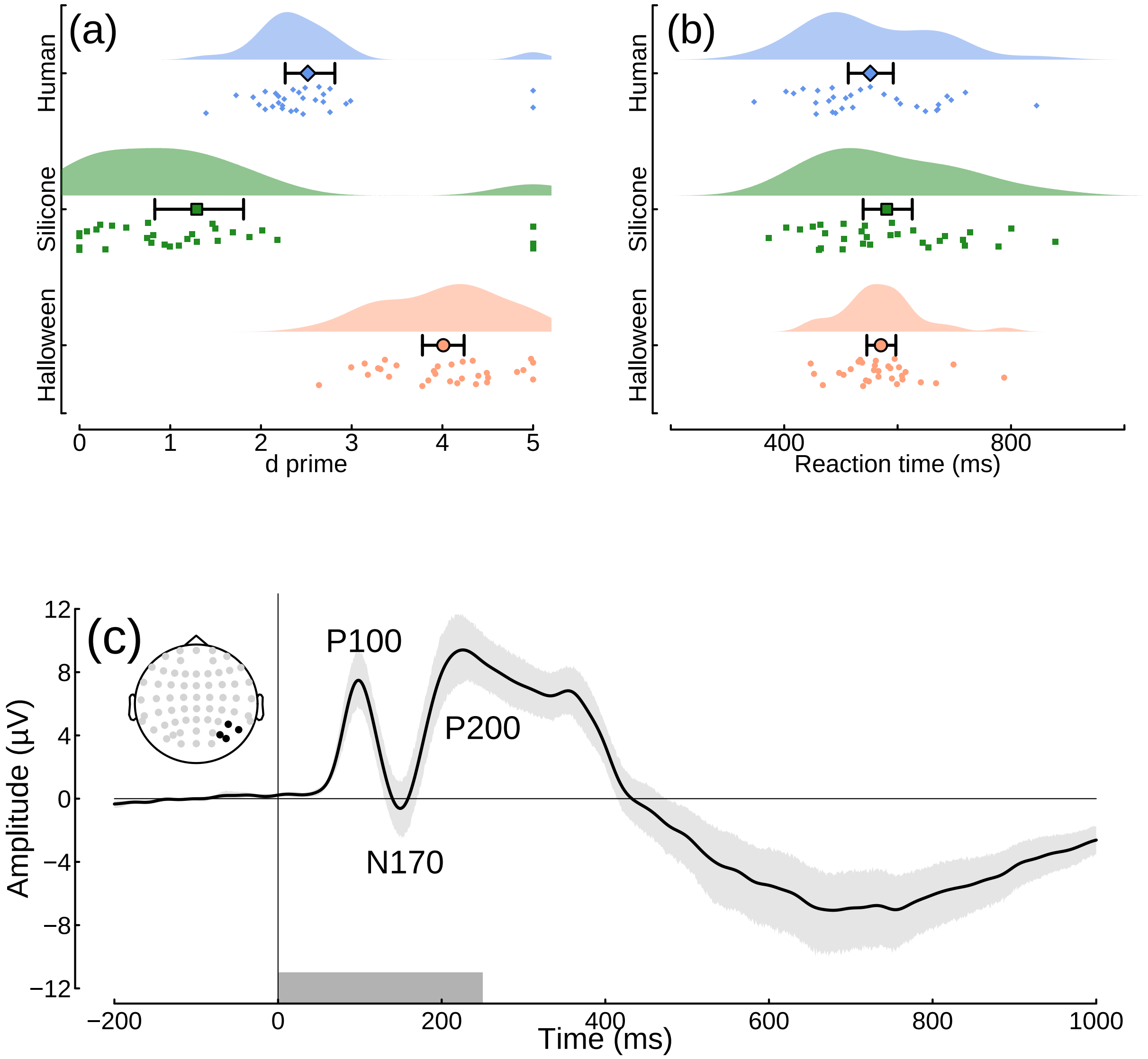

Figure 5: Summary of response data and grand mean ERP for Experiment 2. Panel (a) shows d-prime scores for identifying images of human faces (blue diamonds), silicone masks (green squares) and Halloween masks (red circles). Small points show individual participants, and the larger symbols with error bars indicate the group mean and bootstrapped 95% confidence intervals. Panel (b) plots reaction times in the same format (note the logarithmic x-axis). Panel (c) shows the grand mean ERP across all participants and conditions, pooled across electrodes P6, P8, PO6 and PO8 (see inset). The shaded region around the curve illustrates the 95% confidence interval, and the grey rectangle at the foot indicates the stimulus duration.

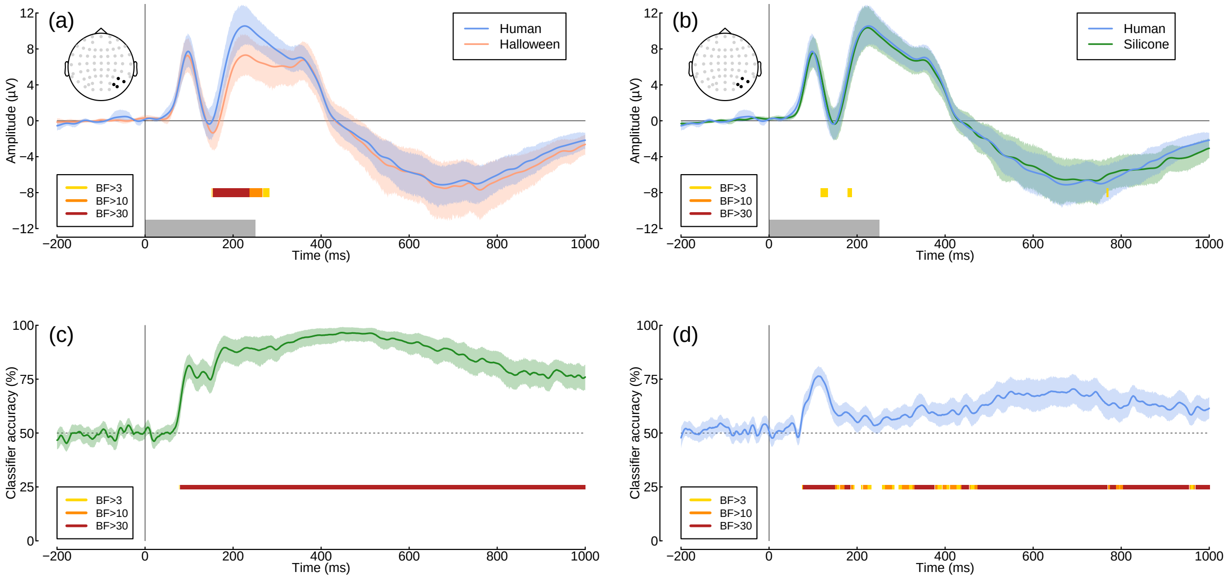

Figure 6: Univariate and multivariate comparisons across image type for Experiment 2. Panel (a) shows the ERPs comparing human faces (blue) and Halloween masks (red), and panel (b) compares human faces (blue) and silicone masks (green). Panels (c) and (d) show multivariate pattern classification accuracy for the same comparisons. Points at y = -8 and y = 25 indicate Bayes factor scores for comparisons between ERPs (a,b) and comparing classification accuracy to chance (50% correct; c,d).

to the unlimited inspection time permitted in this online follow-up experiment

5 Discussion

Across two experiments using diverse stimuli, we identified a potential neurophysiological signature of the 'uncanny valley' effect. EEG responses to androids or silicone masks could be distinguished from responses to human faces at around 100ms after stimulus onset, and also in a later time window around 500-800ms after stimulus onset. There were no clear differences in the unimodal ERP response at posterior electrodes, but performance of a multivariate pattern classifier was above chance in these time windows. This is a different pattern from that observed for more obviously non-human stimuli (robots and Halloween masks), where there were both univariate and multivariate differences, and the multivariate discrimination accuracy was above chance for an extended time window. Perceptual judgements indicated that identification performance for uncanny valley stimuli was relatively poor, indicating confusion with real human images. We also confirmed that android images were perceived more negatively than either human or robot images, and that silicone masks were perceived as more uncanny than human faces. The similarity in results across our two experiments is striking and constitutes an internal conceptual replication of our main findings, suggesting that the neural characteristics of the uncanny valley effect are generalizable across stimulus categories.

The early time window when pattern classification is above chance corresponds approximately to the P100 and N170 components of the ERP. The P100 is typically associated with low-level visual responses, and is affected by contrast and spatial frequency content of an image. The N170 component is most often associated with faces, though is also observed for other image categories, and there is still debate about its precise function [9,10]. Similar early components have also been investigated in other ERP studies on the uncanny valley effect [31,32], and comparing human and robot faces [30]. This time window is unlikely to be modulated substantially by top-down influences, so we attribute the early component to image-based differences between stimulus categories [33]. ERP components in later time windows have also been studied in previous work [17,34,35], and may reflect cognitive processing stages, such as determining whether a stimulus conforms to

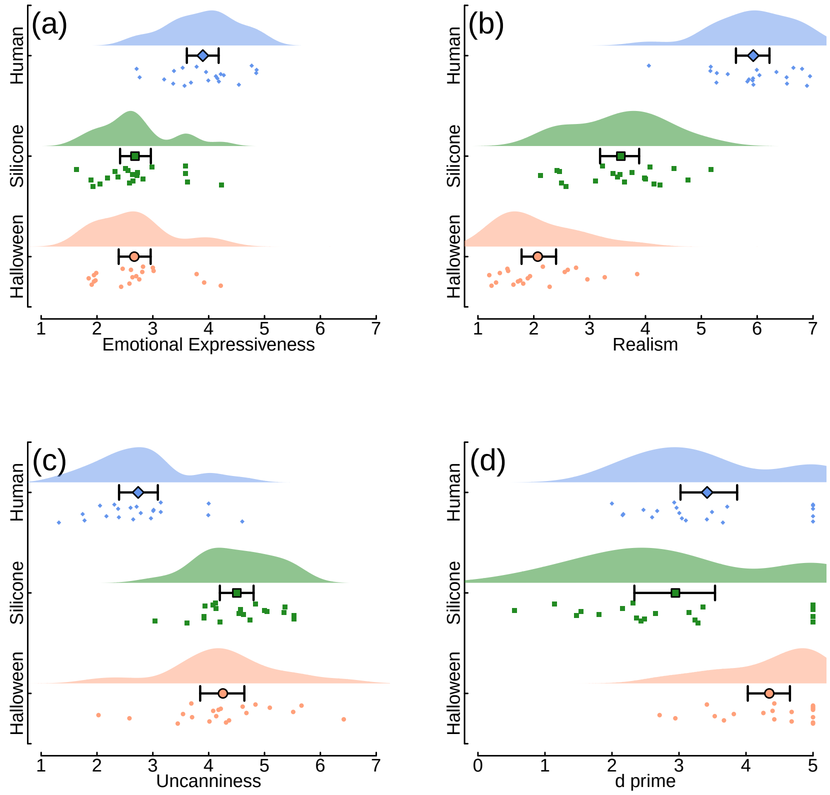

Figure 7: Additional ratings of mask images, completed by an independent sample (N=20) with unlimited inspection time. Images were rated on three dimensions using seven-point scales (panels a-c), and also judged as being either a real face or a mask, from which d-prime scores were calculated (panel d).

categorical expectations. Differences between stimulus categories at these times are more amenable to top-down influences, and most likely involve higher brain areas outside of occipital cortex. We therefore predict that ERP components at later time points should correspond with perceptual judgements and reports of uncanniness - this is a worthwhile hypothesis for future work to investigate.

Our use of hyper-realistic silicone masks is novel in the context of the study of uncanny valley effects. Previous studies using these masks have demonstrated that they are difficult to distinguish from real faces [18,19], including in applied settings such as simulated passport control [36,37], and show large individual differences [38]. Image analysis indicates that good identification performance for silicone masks is typically based on attention to the region below the eyes [38], however it is plausible that many observers are not explicitly aware of the cues they use to perform this judgement. This might contribute to both the early and late components identified in this study, and presumably also to the subjective sensation of 'uncanniness' that is characteristic of the phenomenon.

Another increasingly common situation that triggers the 'uncanny valley' experience is in the domain of computer-generated images and movies [39,40]. Artificial intelligence algorithms are now able to generate images and movies based on text prompts (for example "a picture of a girl flying a kite in a field") that often include human subjects. However, at time of writing, images of humans often contain errors, such as the presence of too many limbs, digits, teeth etc. Synthetic movies often contain continuity errors, and have issues reproducing biological motion. Many of these errors are subtle and take time to spot, but it is also the case that human observers can report that images look 'wrong' without explicitly knowing why. The neural uncanny valley effect that we report here might prove a useful index of these instinctive reactions, and could even potentially be used to improve artificial intelligence algorithms. For example, images could be penalised for producing neural responses that differed from those for natural images.

More generally, the advantage of measuring neural responses to 'uncanny valley' stimuli is that, without requiring conscious awareness or behavioural responses, they can facilitate detection of near-human stimuli. These types of near-human stimuli are becoming increasingly common in impersonation and identity evasion cases [1]. Simultaneously, we observe a growing market for reducing the uncanny valley effect for the benefit of android and robot integration. Exploring the potential of non-invasive brain recordings will benefit various applied fields as a result.

6 Conclusions

We have identified neural correlates of the uncanny valley effect that are consistent across two experiments, using androids and hyper-realistic silicone masks. In both cases, perceptual discrimination from real human faces was possible, but more challenging than discriminating from mechanical robots or Halloween masks. Univariate differences in the ERP signal were unconvincing, but a more sensitive multivariate classification analysis identified differences at both early (100-200ms) and later (around 600ms) time points. These findings suggest the importance of both bottom up and top down influences on the subjective experience of the uncanny valley. Future work might extend these findings to more dynamic stimuli, and explore potential applications for improving android and avatar generation.

References

1. Sanders JG, Jenkins R. Realistic masks in the real world. In: Bindemann M, editor. Forensic Face Matching: Research and Practice. Oxford University Press; 2021. doi:10.1093/oso/9780198837749.003.0010

2. Mori M. The uncanny valley. Energy. 1970;7: 33–35.

3. Mori M, MacDorman KF, Kageki N. The uncanny valley [from the field]. IEEE Robotics & Automation Magazine. 2012;19: 98–100. doi:10.1109/MRA.2012.2192811

4. Vaitonytė J, Alimardani M, Louwerse MM. Scoping review of the neural evidence on the uncanny valley. Computers in Human Behavior Reports. 2023;9: 100263. doi:https://doi.org/10.1016/j.chbr.2022.100263

5. Hu Y, Baragchizadeh A, O'Toole AJ. Integrating faces and bodies: Psychological and neural perspectives on whole person perception. Neurosci Biobehav Rev. 2020;112: 472–486. doi:10.1016/j.neubiorev.2020.02.021

6. Gauthier I, Tarr MJ, Moylan J, Skudlarski P, Gore JC, Anderson AW. The fusiform "face area" is part of a network that processes faces at the individual level. J Cogn Neurosci. 2000;12: 495–504. doi:10.1162/089892900562165

7. Kanwisher N, McDermott J, Chun MM. The fusiform face area: A module in human extrastriate cortex specialized for face perception. J Neurosci. 1997;17: 4302–11. doi:10.1523/JNEUROSCI.17-11-04302.1997

8. Downing PE, Jiang Y, Shuman M, Kanwisher N. A cortical area selective for visual processing of the human body. Science. 2001;293: 2470–3. doi:10.1126/science.1063414

9. Thierry G, Martin CD, Downing P, Pegna AJ. Controlling for interstimulus perceptual variance abolishes N170 face selectivity. Nat Neurosci. 2007;10: 505–11. doi:10.1038/nn1864

10. Hong Y, Mayes MS, Munasinghe AP, Ratner KG. Scrutinizing whether mere group membership influences the N170 response to faces: Results from two preregistered event-related potential studies. J Cogn Neurosci. 2022;34: 1999–2015. doi:10.1162/jocn_a_01887

11. MacDorman KF. Does mind perception explain the uncanny valley? A meta-regression analysis and (de)humanization experiment. Computers in Human Behavior: Artificial Humans. 2024;2: 100065. doi:https://doi.org/10.1016/j.chbah.2024.100065

12. MacDorman KF, Chattopadhyay D. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition. 2016;146: 190–205. doi:10.1016/j.cognition.2015.09.019

13. Moore RK. A bayesian explanation of the 'uncanny valley' effect and related psychological phenomena. Sci Rep. 2012;2: 864. doi:10.1038/srep00864

14. Gray K, Wegner DM. Feeling robots and human zombies: Mind perception and the uncanny valley. Cognition. 2012;125: 125–30. doi:10.1016/j.cognition.2012.06.007

15. Yam KC, Bigman Y, Gray K. Reducing the uncanny valley by dehumanizing humanoid robots. Computers in Human Behavior. 2021;125: 106945. doi:https://doi.org/10.1016/j.chb.2021.106945

16. Saygin AP, Chaminade T, Ishiguro H, Driver J, Frith C. The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Soc Cogn Affect Neurosci. 2012;7: 413–22. doi:10.1093/scan/nsr025

17. Urgen BA, Kutas M, Saygin AP. Uncanny valley as a window into predictive processing in the social brain. Neuropsychologia. 2018;114: 181–185. doi:10.1016/j.neuropsychologia.2018.04.027

18. Sanders JG, Ueda Y, Minemoto K, Noyes E, Yoshikawa S, Jenkins R. Hyper-realistic face masks: A new challenge in person identification. Cognitive Research: Principles and Implications. 2017;2. doi:10.1186/s41235-017-0079-y

19. Sanders JG, Ueda Y, Yoshikawa S, Jenkins R. More human than human: A Turing test for photographed faces. Cogn Res Princ Implic. 2019;4: 43. doi:10.1186/s41235-019-0197-9

20. Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics. 2009;1: 71–81. doi:10.1007/s12369-008-0001-3

21. Koverola M, Kunnari A, Sundvall J, Laakasuo M. General attitudes towards robots scale (GAToRS): A new instrument for social surveys. International Journal of Social Robotics. 2022;14: 1559–1581. doi:10.1007/s12369-022-00880-3

22. Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J Autism Dev Disord. 2001;31: 5–17. doi:10.1023/a:1005653411471

23. Macmillan NA, Creelman CD. Detection theory: A user's guide. Psychology Press; 2005.

24. Delorme A, Makeig S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004;134: 9–21. doi:10.1016/j.jneumeth.2003.10.009

25. Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM. Brainstorm: A user-friendly application for MEG/EEG analysis. Comput Intell Neurosci. 2011;2011: 879716. doi:10.1155/2011/879716

26. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev. 2009;16: 225–37. doi:10.3758/PBR.16.2.225

27. Jeffreys H. Theory of probability. 3rd ed. Oxford University Press, Clarendon Press; 1961.

28. Chang C-C, Lin C-J. LIBSVM : A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2(27): 1–27.

29. Grootswagers T, Wardle SG, Carlson TA. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. J Cogn Neurosci. 2017;29: 677–697. doi:10.1162/jocn_a_01068

30. Geiger AR, Balas B. Robot faces elicit responses intermediate to human faces and objects at face-sensitive ERP components. Sci Rep. 2021;11: 17890. doi:10.1038/s41598-021-97527-6

31. Schindler S, Zell E, Botsch M, Kissler J. Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. Sci Rep. 2017;7: 45003. doi:10.1038/srep45003

32. Mustafa M, Magnor M. EEG based analysis of the perception of computer-generated faces. Proceedings of the 13th european conference on visual media production (CVMP 2016). ACM; 2016. doi:10.1145/2998559.2998563

33. Coggan DD, Baker DH, Andrews TJ. The role of visual and semantic properties in the emergence of category-specific patterns of neural response in the human brain. eNeuro. 2016;3. doi:10.1523/ENEURO.0158-16.2016

34. Mustafa M, Guthe S, Tauscher J-P, Goesele M, Magnor M. How human am i?: EEG-based evaluation of virtual characters. Proceedings of the 2017 CHI conference on human factors in computing systems. ACM; 2017. doi:10.1145/3025453.3026043

35. Cheetham M, Wu L, Pauli P, Jancke L. Arousal, valence, and the uncanny valley: Psychophysiological and self-report findings. Front Psychol. 2015;6: 981. doi:10.3389/fpsyg.2015.00981

36. Robertson DJ, Sanders JG, Towler A, Kramer RSS, Spowage J, Byrne A, et al. Hyper-realistic face masks in a live passport-checking task. Perception. 2020;49: 298–309. doi:10.1177/0301006620904614

37. Robertson DJ, Davis JP, Sanders JG, Towler A. The super-recogniser advantage extends to the detection of hyper-realistic face masks. Applied Cognitive Psychology. 2024;38: e4222. doi:10.1002/acp.4222

38. Sanders JG, Jenkins R. Individual differences in hyper-realistic mask detection. Cogn Res Princ Implic. 2018;3: 24. doi:10.1186/s41235-018-0118-3

39. Moshel ML, Robinson AK, Carlson TA, Grootswagers T. Are you for real? Decoding realistic AI-generated faces from neural activity. Vision Res. 2022;199: 108079. doi:10.1016/j.visres.2022.108079

40. Gu Z, Jamison K, Sabuncu MR, Kuceyeski A. Human brain responses are modulated when exposed to optimized natural images or synthetically generated images. Commun Biol. 2023;6: 1076. doi:10.1038/s42003-023-05440-7