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

Shona Fitzpatrick, Ailish K. Byrne, Alex Headley, Jet G. Sanders,
Helen Petrie, Rob Jenkins & Daniel H. Baker (daniel.baker@york.ac.uk)

University of York

1 Abstract

2 Introduction

Many people report an aversion to entities that are superficially human-like, but on closer inspection are actually artificial. Examples include humanoid robots (androids), puppets, hyper-realistic masks, and computer-generated images or movies. The term 'uncanny valley' (Mori, 1970; English translation in Mori et al., 2012) describes the idea that entities that are clearly human or clearly artificial 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 Vaitonytė et al., 2023).

Neural responses to faces and bodies in general are well-characterised, and there appear to be specialised brain regions devoted to both (reviewed in Hu et al., 2020). For example, areas of the occipital lobe (Gauthier et al., 2000) and fusiform gyrus (Kanwisher et al., 1997) respond more to faces than non-face stimuli, and sections of extrastriate cortex are responsive to bodies (Downing et al., 2001). There are also event-related potential (ERP) signals associated with face and body stimuli, though their precise purpose is still debated (Thierry et al., 2007). It seems highly likely that 'uncanny' images will activate these same processes, yet it is unclear whether the sense of unease they produce occurs at bottom-up sensory stages, or is modulated by more top-down cognitive factors.

A previous study by Saygin et al. (2012) measured fMRI responses to moving stimuli designed to elicit an uncanny valley effect. They found repetition suppression effects 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 (Urgen et al., 2018) identified a difference in the N400 component 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 temporo-parietal cortex, consistent with the fMRI results (Saygin et al., 2012). 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 here was to further investigate neural correlates of the uncanny valley effect. We achieve this through two EEG experiments, in which we measure neural responses to static images. In the first experiment, the stimuli were 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. carnival or Halloween masks), and wearing hyper-realistic silicone masks (Sanders et al., 2017) (see Figure 1b). Rather than focus on specific ERP components, we use a multivariate pattern classification approach 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.

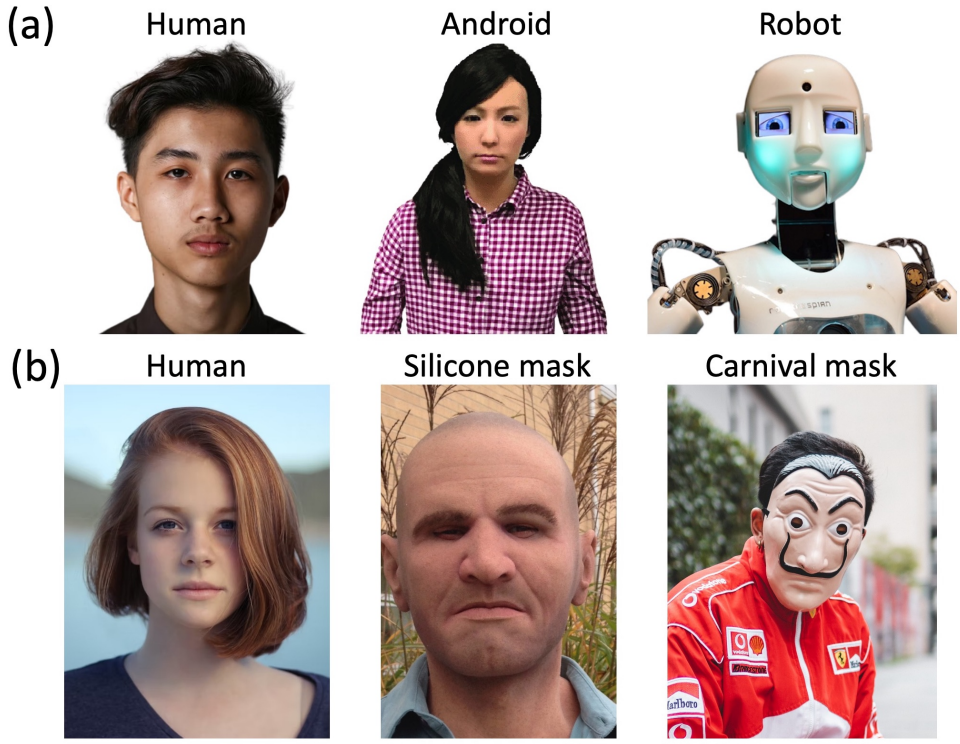

Figure 1: Illustrative stimuli from the same categories as used in Experiments 1 and 2. Row (a) shows a human face, an android and a robot, all against white backgrounds. Row (b) shows a human face, a hyper-realistic silicone mask, and a carnival mask, all against natural backgrounds. Images were taken 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, and was used in Experiment 2.

3 Materials & Methods

3.1 Participants

A total of 29 participants completed Experiment 1 (12 male, 17 female), and 30 participants completed Experiment 2 (7 male, 23 female). 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.

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 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. In both experiments, images involved 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. 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. 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 6 minutes.

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 (Sanders et al., 2017), however as that is not the main focus of the current paper 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.

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 (Macmillan and Creelman, 2005). 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. 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 (Delorme and Makeig, 2004) to import the data into Matlab and collate data across blocks. We then used Brainstorm (Tadel et al., 2011) 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 was unsuccessful. Their behavioural data were still included in the analysis. In Experiment 2 we combined data across all three image sizes, as this was not the main focus of our analysis.

We performed univariate analyses by conducting Bayesian t-tests (Rouder et al., 2009) between ERPs from pairs of conditions at each time point using a JZS prior. 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 (1961) 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 was conducted by training a linear support vector machine algorithm (LibSVM; Chang and Lin (2011)) to discriminate between patterns of activity across electrodes at a specific time point. 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 using these to train the classifier. The accuracy of the classifier was tested on the averaged 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 also for all participants. 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://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.

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 (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, however the evidence for a convincing difference 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.

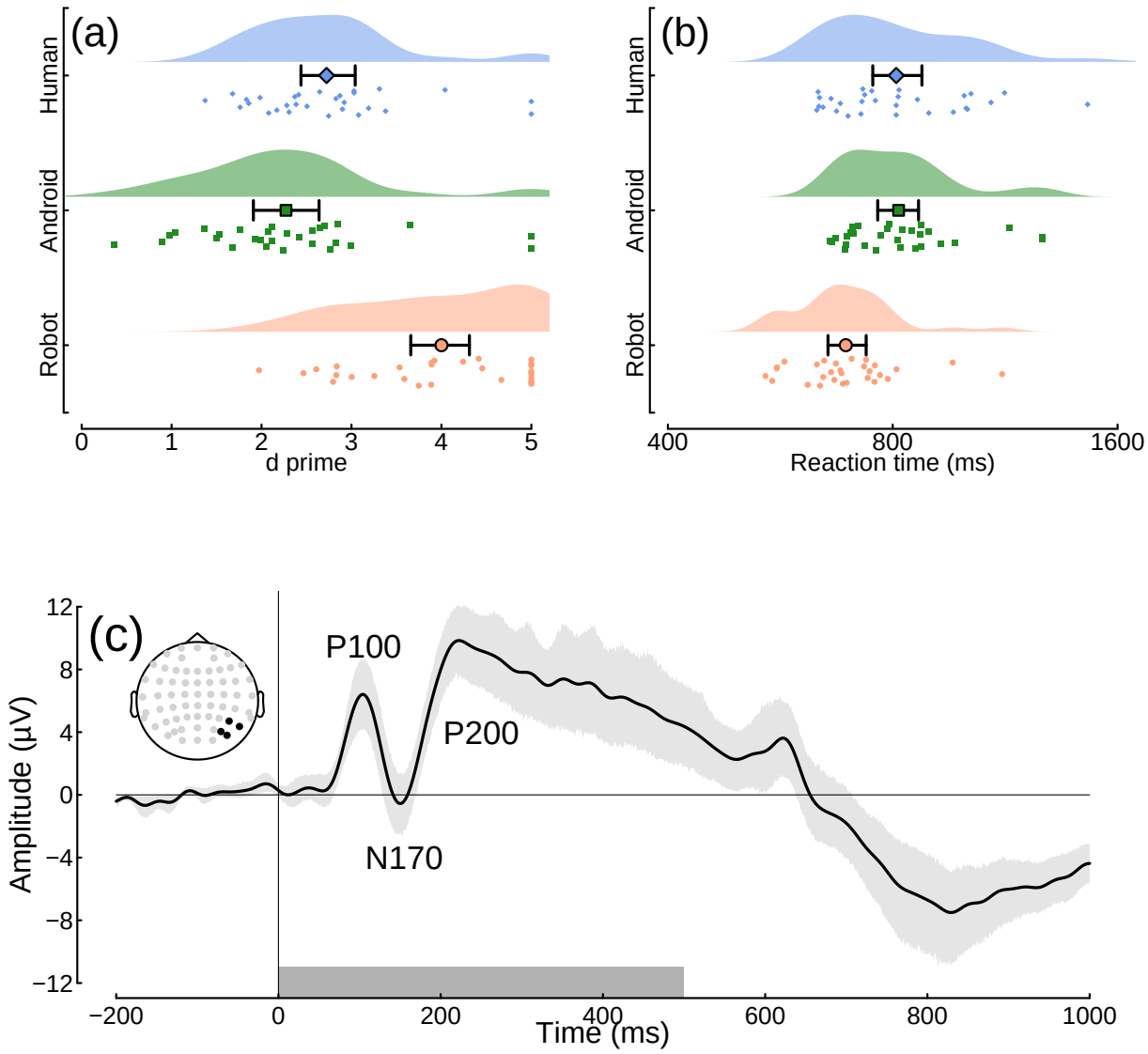

Figure 2: Summary of response data and grand mean ERP for Experiment 1. Panel (a) shows d-prime scores for identifying images of human (blue), android (green) and robot (red) faces. Small points show individual participants, and the larger symbols with error bars indicate the group mean and bootstrapped 95 percent 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 percent confidence interval, and the grey rectangle at the foot indicates the stimulus duration.

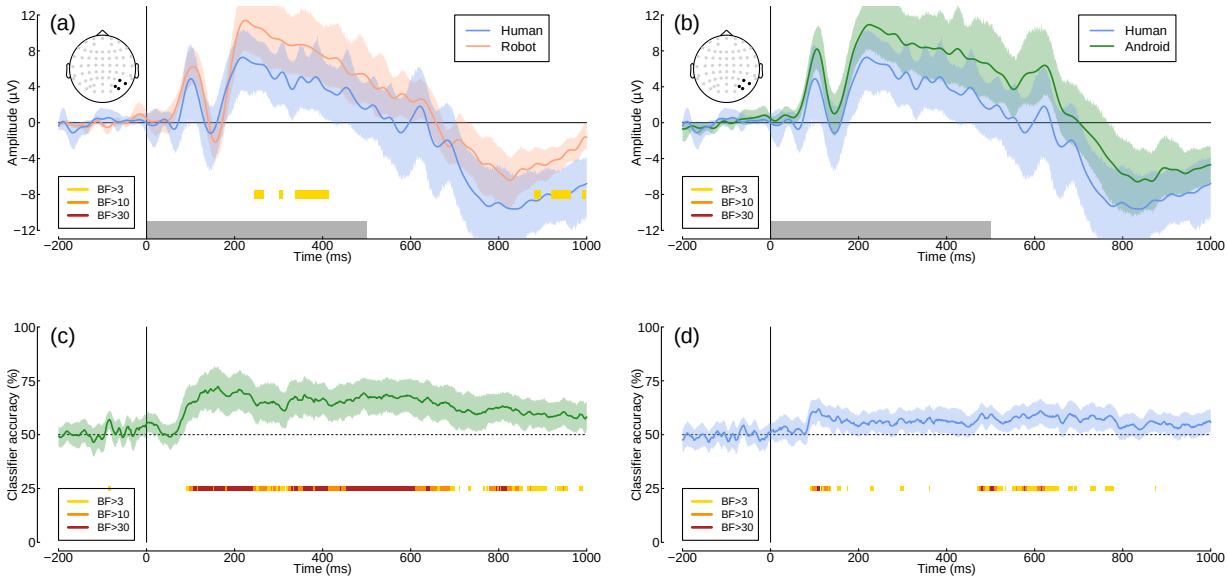

Figure 3: Univariate and multivariate comparisons across image type. Panel (a) shows the ERPs comparing human (blue) and robot (red) face images, and panel (b) compares human (blue) and android (green) faces. Panels (c) and (d) show multivariate pattern classification accuracy for the same comparisons.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

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 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. In the Discussion we speculate that these two time periods might correspond to distinct types of signal associated with the uncanny valley. However we first 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 4a. Pairwise Bayesian t-tests between conditions indicate very convincing differences in sensitivity between 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 4b.

The grand average ERP waveform for Experiment 2 (see Figure 4c) had similar initial components as for Experiment 1. The latter portion of the waveforms differed somewhat, most likely owing to the difference

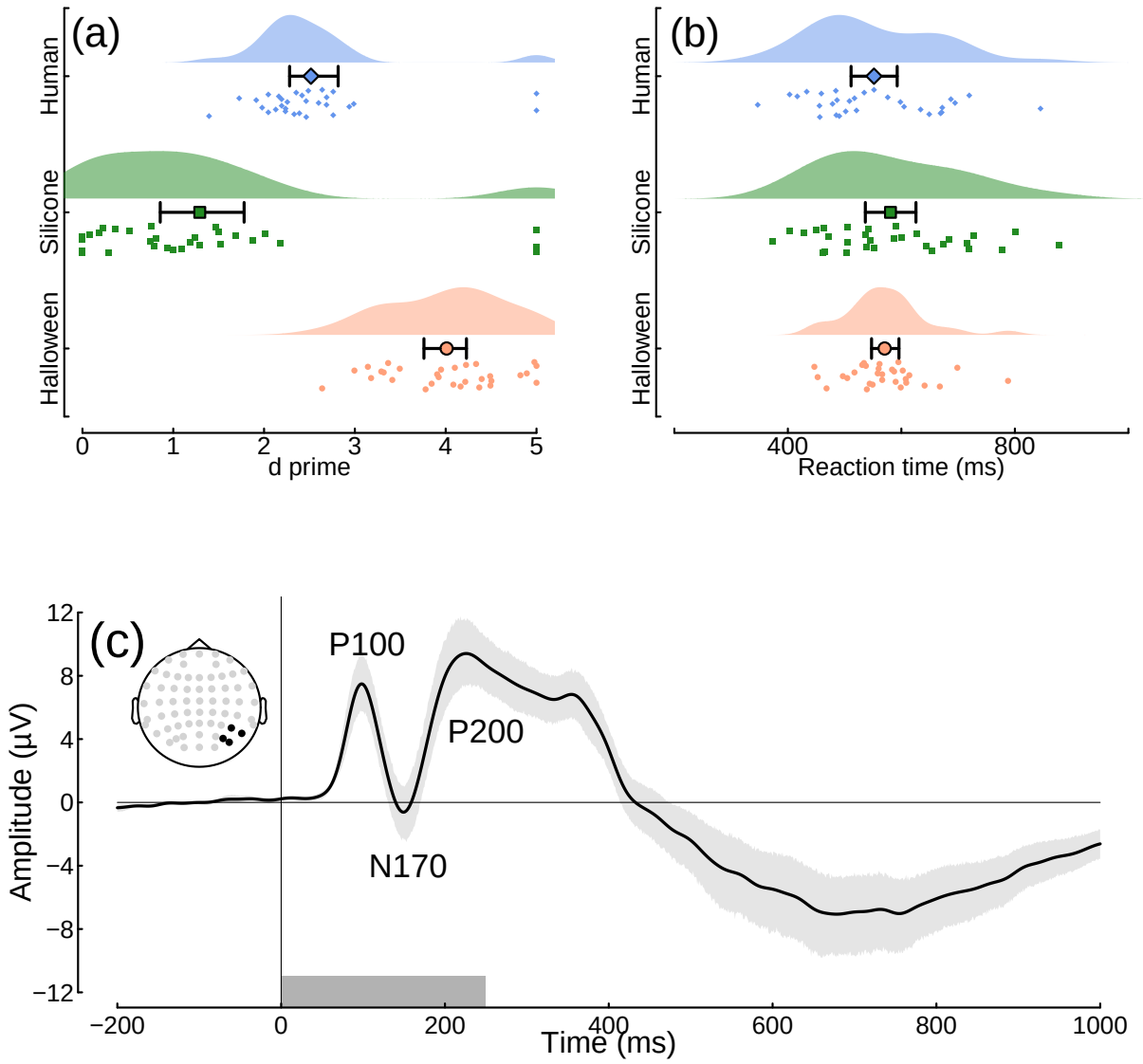

Figure 4: Summary of response data and grand mean ERP for Experiment 2. Panel (a) shows d-prime scores for identifying images of human faces (blue), silicone masks (green) and Halloween masks (red). Small points show individual participants, and the larger symbols with error bars indicate the group mean and bootstrapped 95 percent 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 percent confidence interval, and the grey rectangle at the foot indicates the stimulus duration.

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 5a), with Bayes factors exceeding 30. Univariate differences between the human and silicone mask conditions were not compelling (see Figure 5b).

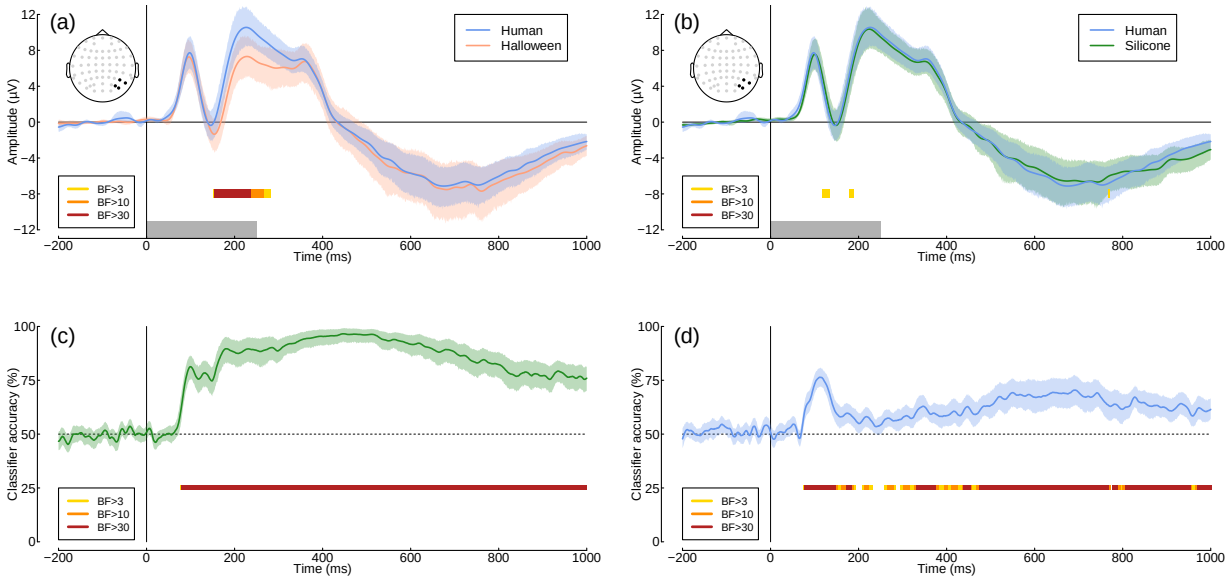

Figure 5: Univariate and multivariate comparisons across image type for Experiment 2. Panel (a) shows the ERPs comparing human faces (blue) and Halloween masks (red), and panel (b) compares human faces (blue) and silicone masks (green). Panels (c) and (d) show multivariate pattern classification accuracy 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 percent correct; c,d).

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 5c). Classification was also convincingly above chance when comparing human faces with silicone masks (Figure 5d). 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.

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. The similarity in results across our two experiments is striking, and constitutes an internal conceptual replication of our main findings.

Big theory

Another increasingly common situation that triggers the 'uncanny valley' experience is in the domain of computer-generated images and movies. 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 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.

6 Conclusions

7 Acknowledgements

SF was funded by a YorRobots Venables internship. Also supported by BBSRC grant BB/V007580/1 awarded to DHB. ANY OTHER GRANTS TO ACKNOWLEDGE HERE?

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Vaitonytė J, Alimardani M, Louwerse MM. 2023. Scoping review of the neural evidence on the uncanny

valley. *Computers in Human Behavior Reports* **9**:100263. doi:https://doi.org/10.1016/j.chbr.2022.100263