Neural correlates of the uncanny valley effect for robots and hyper-realistic masks (for submission to Computers in Human Behavior)

Shona Fitzpatrick, Ailish K. Byrne, Alex Headley, Jet G. Sanders,
Helen Petrie, Rob Jenkins & Daniel H. Baker

2023-01-02

1 Abstract

2 Introduction

Many people report an aversion to entities that are human-like, but on closer inspection are actually artificial. Examples include humanoid robots (androids), puppets, realistic computer-generated images, and hyper-realistic masks. 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.

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.

One previous study has measured fMRI responses to moving stimuli designed to elicit an uncanny valley effect. Saygin et al. (2012) 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. 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). Rather than focus on

specific ERP components, we use a non-parametric cluster correction procedure to compare conditions. We also apply a pattern classification approach to identify time windows in which information in the EEG signal can be used to distinguish between pairs of conditions.

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 Halloween masks (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. 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

EEG signals were recorded during each block, and saved to disc for subsequent offline analysis. We used a component of the EEGlab toolbox (Delorme and Makeig, 2004) to convert the continuous data from a proprietary file format to a compressed csv text file. All subsequent analyses were conducted in R using these files.

For each block, data at each electrode were low-pass filtered at 30Hz with a 10^{th} order Butterworth filter, and then epoched using the stimulus onset triggers. A pre-stimulus baseline (average voltage of the 200ms before stimulus onset) was subtracted from each waveform. We rejected ERPs that showed evidence of excessive noise or movement artefacts on a per-electrode basis by excluding trials where the standard deviation across the time window from 200ms before to 1000ms after stimulus onset exceeded $40\mu V$. ERPs were averaged across trials for each participant, and then across participants to calculate group averages.

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 constitute some evidence supporting the alternative hypothesis, factors >10 constitute strong evidence, and factors >30 constitute very strong evidence.

Multivariate pattern analysis was conducted by training a linear support vector machine algorithm to discriminate between patterns of activity across electrodes at a specific time point. The patterns came from the real face condition and one of the mask conditions at a single stimulus size, and for a single participant. Examples of each pattern were calculated by averaging over random subsets of 36 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 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/

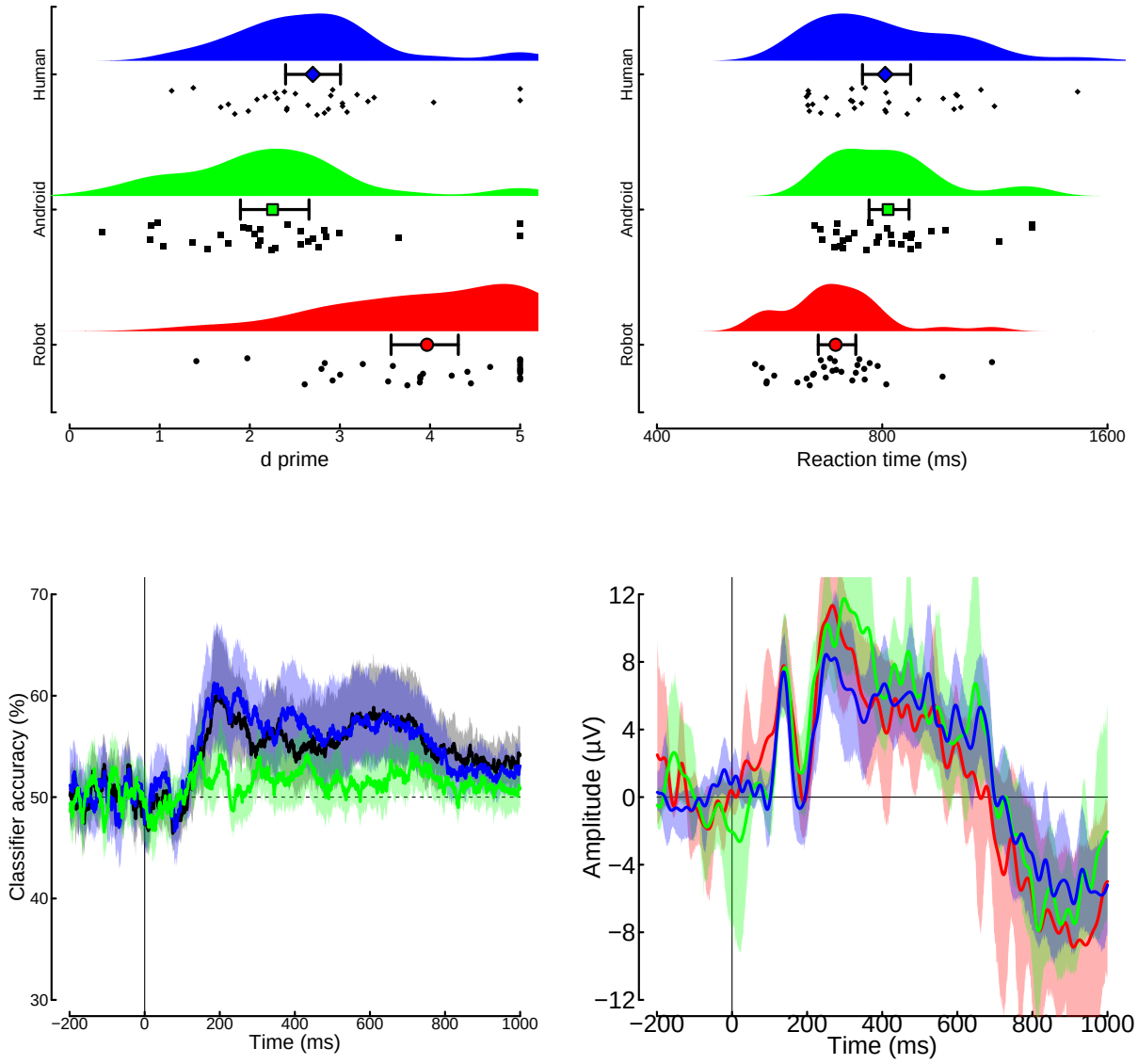
Human
Android
Robot
0
1
2
3
4
5
d prime
400
800
1600
Reaction time (ms)
70
60
50
40
30
Classifier accuracy (%)
−200
0
200
400
600
800
1000
Time (ms)
12
8
4
0
−4
−8
−12
Amplitude (μV)
Time (ms)

Figure 1: I.

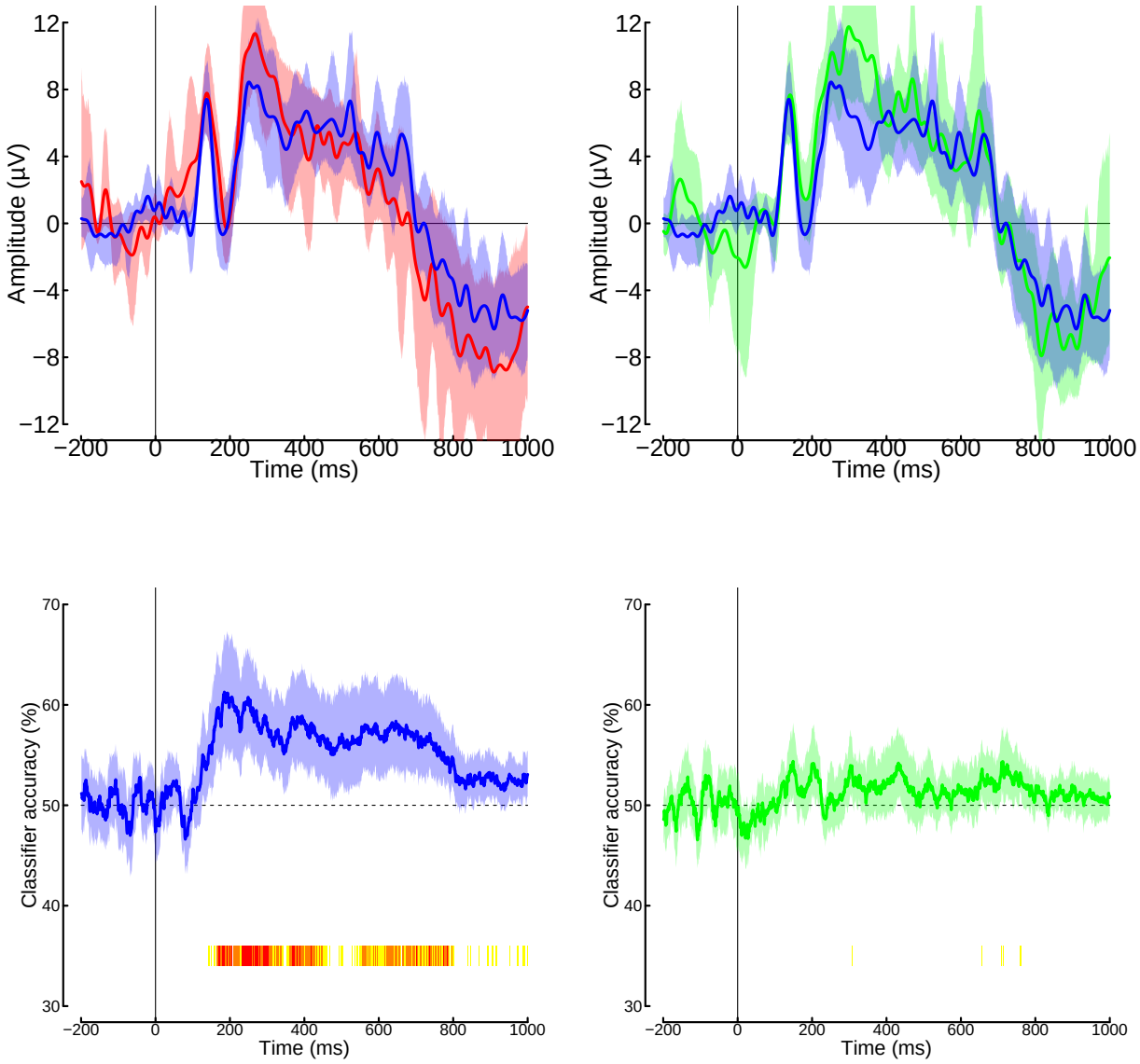

Figure 2: I.

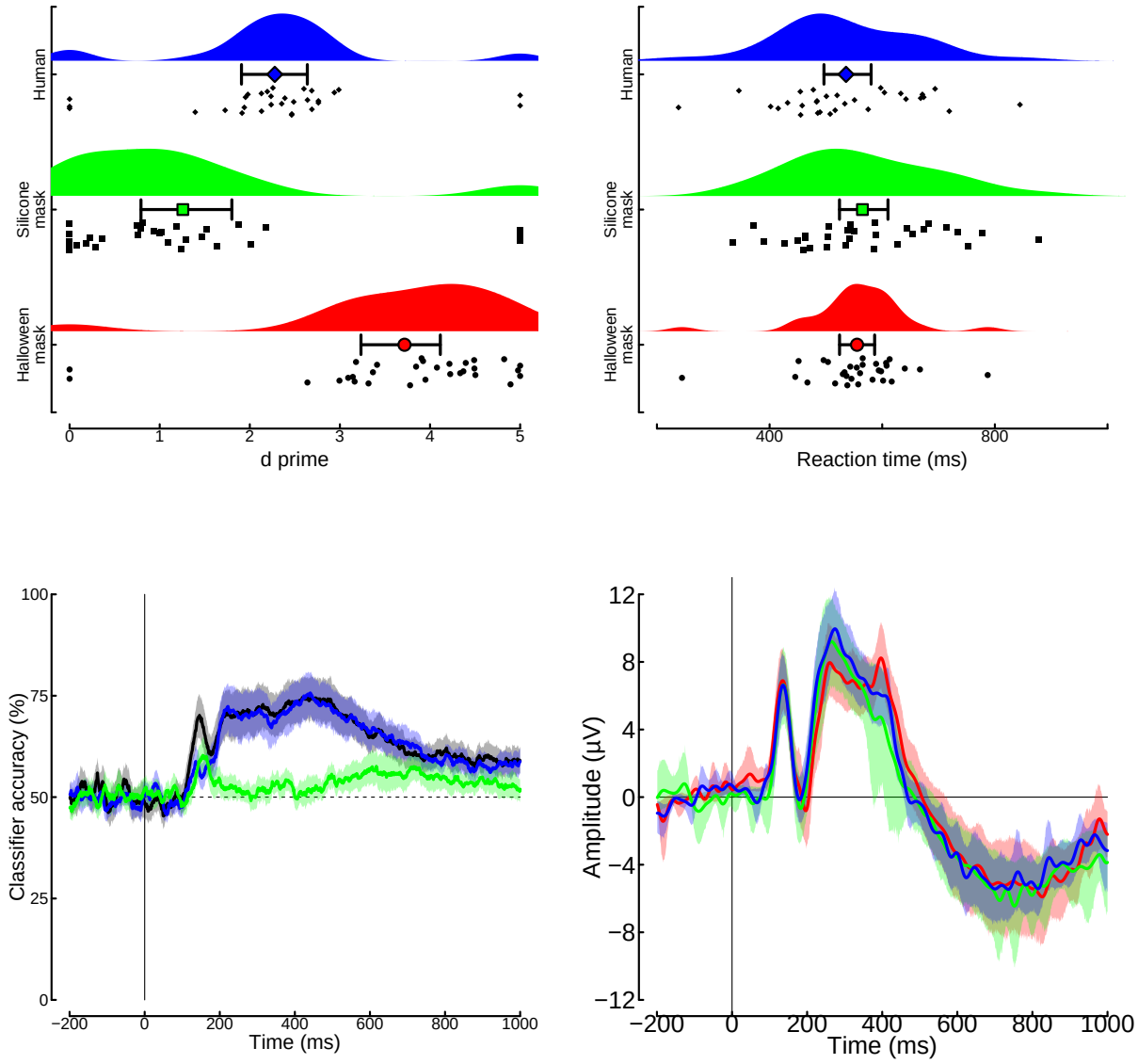
Human
Silicone mask
Halloween mask
0
1
2
3
4
5
d prime
400
800
Reaction time (ms)
100
75
50
25
0
Classifier accuracy (%)
−200
0
200
400
600
800
1000
Time (ms)
12
8
4
0
−4
−8
−12
Amplitude (μV)
−200
0
200
400
600
800
1000
Time (ms)

Figure 3: I.

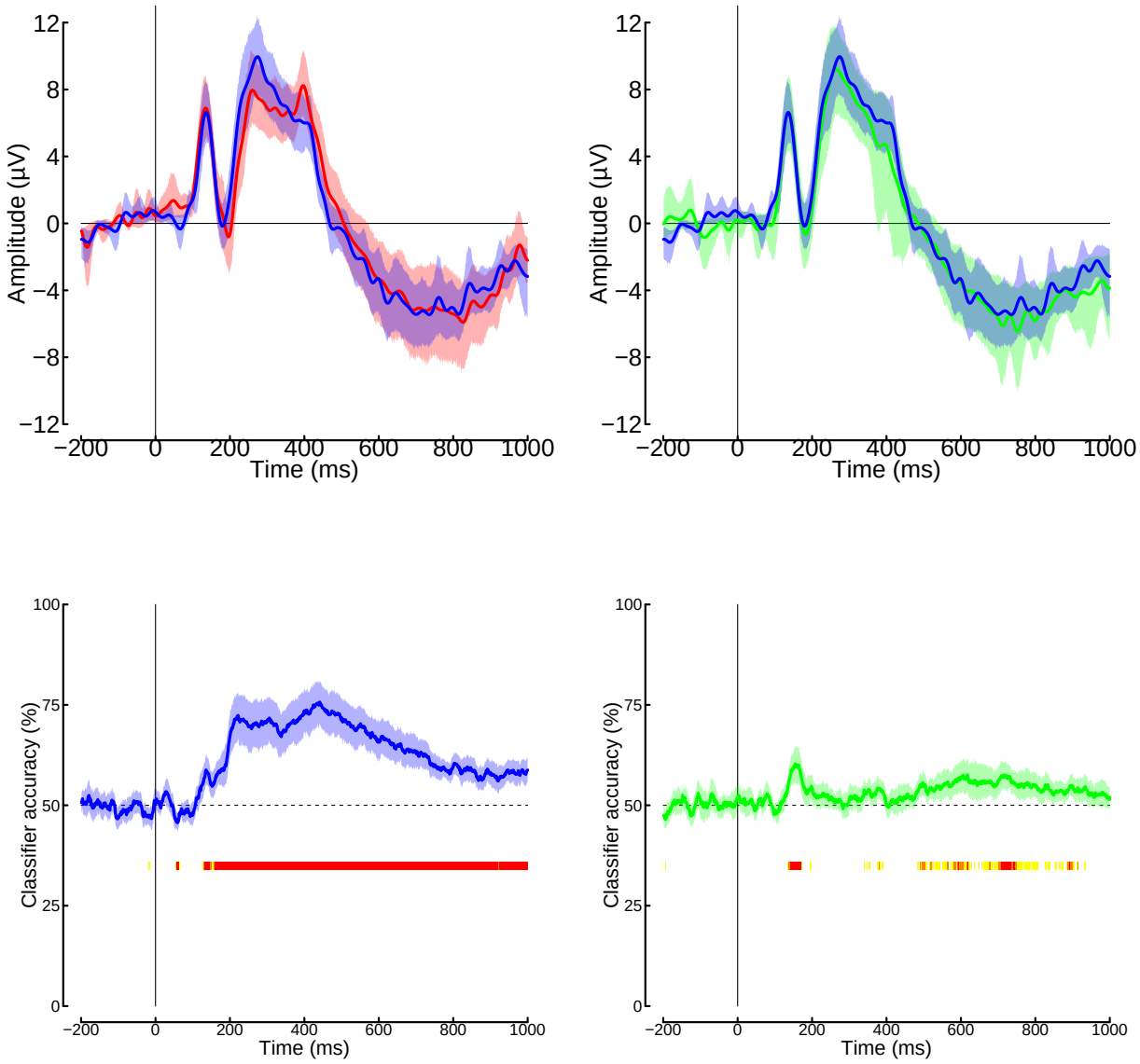

Figure 4: I.

4 Results

4.1 Experiment 1

4.2 Experiment 2

5 Discussion

6 Conclusions

7 Acknowledgements

References

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

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

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