Are you for real? Decoding realistic AI-generated faces from neural activity

Michoel L. Moshel [a,b]*, Amanda K. Robinson[a], Thomas A. Carlson [a#], Tijl Grootswagers[a,c#]

[a] School of Psychology, University of Sydney, NSW, Australia

[b] School of Psychology, Macquarie University, NSW, Australia

[c] The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, NSW, Australia

[#] shared authorship

* corresponding author: michoel.moshel@hdr.mq.edu.au

Significance Statement

The generation of realistic images and faces has become increasingly possible in the last few years due to advances in artificial intelligence (AI) technologies. Given the human expertise and specialization for face perception, understanding how the brain is fooled by realistic AI-generated faces may be crucial in navigating this new era of realism. We investigated how humans deal with realistic faces using behavioural data and computational neuroimaging. We found that whilst we could reliably decode AI-generated faces using people's neural activity, observers could not reliably discriminate between real and realistic fake faces. Understanding this dissociation has significant implications for fake face detection as well as face perception in general.

Abstract

Can we trust our eyes? Until recently, we rarely had to question whether what we see is indeed what exists, but this is changing. Artificial neural networks can now generate realistic images that challenge our perception of what is real. This new reality can have significant implications in cybersecurity, counterfeiting, fake news, and border security. We investigated how the human brain encodes and interprets realistic artificially generated images using behaviour and brain imaging. We found that we could reliably detect AI-generated fake images using neural activity, even though people had difficulty telling apart real faces from realistic fakes. Understanding this difference between brain and behavioural responses may be key in determining the 'real' in our new reality. Stimuli, code, and data for this study can be found at https://osf.io/n2z73/.

Keywords

Face perception Decoding Fake faces Artificial intelligence Neuroimaging

Introduction

The novel and rapidly emerging phenomena of fake multimedia have swept through modern culture to the extent that the fake has become the expected norm (Adelani et al., 2020; Shen et al., 2019; Shu et al., 2017). The degree to which terms like 'fake news' or 'photoshopped' have become common parlance is indicative of a general and commonly experienced inability to distinguish between what is real and what is not (Fletcher, 2018). Meanwhile, AI technologies, in particular Generative Adversarial Networks (GANs), have been making increasingly rapid advances in generating realistic images with face generation as a major focus (Karras et al., 2019, 2020; Wang et al., 2018; Yu et al., 2020). These advances in realism have begun to have real-world consequences including undetectable videos of fake events ("Deepfakes": Kietzmann et al., 2020), art and audio-visual counterfeits (Farokhmanesh, 2018), and fraudulent social media accounts (Gleicher, 2019). For instance, in 2019, Facebook announced that fake accounts were being created with profile pictures generated by artificial intelligence in an attempt to evade detection (Gleicher, 2019). Crucially, understanding how people respond to AI images, in terms of both behaviour and neural responses, will inform us about how realistic artificial images and faces are perceived differently to real ones, how this dissociation is encoded by the brain, and can ultimately aid in the development of future policy and strategies to curb the potentially nefarious uses of fake media.

One area in which AI technology has made increasingly rapid and apparent progress in is the generation of realistic faces. Until now, fooling observers with artificial faces has been a particularly difficult given the expertise humans have with face perception and recognition (Farid & Bravo, 2007, 2012; Gauthier & Tarr, 2002; Sinha et al., 2006). Not only are faces perceived differently than objects (Shakeshaft & Plomin, 2015; Sunday et al., 2019) but neuroimaging studies highlight distinct brain networks for face processing (Axelrod & Yovel, 2015; Gauthier & Tarr, 2002). The specialized processing of faces results in the rapid and automatic detection of artificial face appearance (Wheatley et al., 2011). For example, the uncanny valley effect describes how observers remain viscerally aware of artificial faces indicated by a steady drop in affinity as an artificial face approaches human likeness, despite not being able to identify any perceivable defects (MacDorman & Chattopadhyay, 2016). In another example, photographs of real faces yield a higher recognition accuracy than computer-generated equivalents demonstrative of enhanced face expertise for the former (Crookes et al., 2015). Likewise, observers have typically performed well at discriminating human faces from computer-generated faces depending on image resolution, training, and incentives (Holmes et al., 2016). However, more recent studies have shown increasingly poorer performances at telling real from fake (Mader et al., 2017; Nightingale et al., 2017; Sanders et al., 2019; Zhou et al., 2019). As the capacity for image realism is steadily increasing, identification of fake faces will likely be further challenged.

Neuroimaging has provided useful insight into how face perception unfolds over time. Electroencephalography (EEG), which measures electrical activity at the scalp with very high temporal resolution, has been used to identify unique neural responses that reflect the temporal emergence and dynamics of facial processing (Bentin et al., 1996; Rossion et al., 2000). Wheatley and colleagues (2011) demonstrated the brain's discrimination of real and artificial faces by comparing neural responses to real faces with responses to doll faces. The authors found that both human and artificial faces elicited an N170, a face-specific neural response approximately 170ms after image presentation. However, sustained positivity beyond 400ms was associated only with human faces, suggesting that this EEG potential could index a process that distinguishes between real and fake faces (Wheatley et al., 2011). Indeed, in other studies, sustained positivity, characterised by the late positive amplitude (LPP), increased as face realism increased, suggesting that real faces, more so than artificial faces, engage high-level attentional, semantic and identity evaluations (Schindler et al., 2017). The new generation of realistic faces produced by GAN technology, however, is of a far superior quality than previously studied artificial faces and often practically indistinguishable from real faces. Whether the brain elicits neural indicators consistent with artificial fake detection for the new generation of GAN-produced images has yet to be seen. Considering that humans remain the gold standard of fake image and face detection (Natsume et al., 2019, Marra et al., 2018), examining the neural mechanisms in fake face detection is instrumental in understanding how to best tackle and understand the new age of fake media. EEG remains an ideal method to provide useful insights into the neural processing of fake GAN faces. Firstly, it allows for an insight into the sequential stages of face processing, from low-level visual features to holistic face perception. Secondly, closer examination at the neuronal population level enables us to answer at what temporal stages GAN face perception may differ from real face perception. Thirdly, using newer multivariate methods applied to EEG data enables analysis of signal-level information on a trial-by-trial basis and can pinpoint the precise temporal emergence of visual processing (Grootswagers, Robinson, & Carlson, 2019; Haynes & Rees, 2006; Teichmann et al., 2020).

With progressive advances in realistic image generation, have we reached a point where observers can no longer tell apart real from the fake? Can measuring the brain's response reveal how realistic fake faces are distinguished from real faces? We measured whether observers could behaviourally discriminate real faces from GAN-generated faces at two levels of face realism; one level of realism similar to fake images used in previous work ("unrealistic"), and another level which represents the current state-of-the-art realistic artificial images ("realistic"). We expected that participants would not be able to discriminate real from realistic faces but could for unrealistic faces, consistent with previous research using AI-generated faces (Hulzebosch et al., 2020; Zhou et al., 2019). To investigate whether we could decode real and fake images from brain activity we used time-resolved multivariate pattern analysis (MVPA) and EEG. To ensure the real and fake stimuli evoked typical categorical effects that could be decoded in the neural signal, we also included cars and bedrooms stimuli. We presented images upright in rapid sequences, which we have previously shown captures low- and high-level image processing (Grootswagers, Robinson, & Carlson; Oosterhof et al., 2016). To determine the contribution of low-level image properties, we used a much faster presentation rate (20Hz; Robinson et al., 2019) and also investigated how real/fake face processing is affected by image inversion, which limits high-level expert face-processing. Consistent with the brain's sensitivity to artificial face appearance, we found it was possible to decode real faces from GAN-generated faces at both levels of face realism using the EEG data. When asked to behaviourally classify faces as either real or fake, a large group of participants could correctly classify or spot both unrealistic and realistic fake faces. However, they performed below chance at classifying real faces from realistic faces. In other words, observers perceived GAN realistic faces as appearing more real than real faces. Understanding differences between observer-reported perceptions of fake images and the brain's response can yield important insights into human face perception in general as well as raise possibilities for training observers to tell apart real from fake.

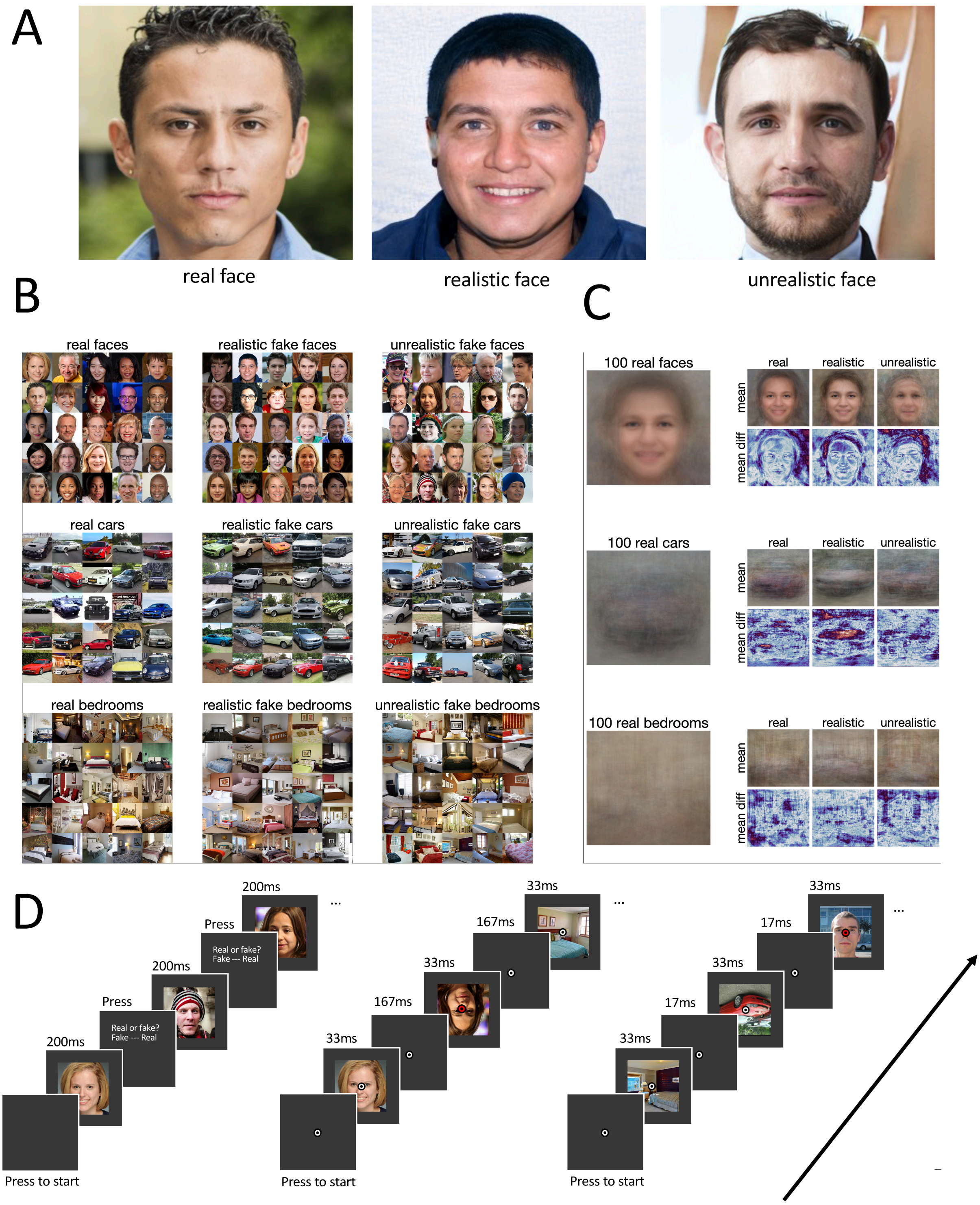

Figure 1. Experimental stimuli and design. A) Real, realistic, and unrealistic faces used in the experiment, B) Face, car, and bedroom stimuli from three conditions (real, realistic fake, unrealistic fake), taken from StyleGAN. C) Mean image for each condition and the absolute pixel difference between 100 independent real images not used in the experiment. Brighter colours (orange) indicate greater absolute differences. D) Experimental designs from left to right; behavioural experiment, 5Hz EEG experiment and 20Hz EEG experiment.

Methods

We performed two experiments that investigated fake versus real image identification: one behavioural and one neuroimaging. The stimuli, data, and analysis code can be found at https://osf.io/n2z73/.

Participants

For behavioural testing, we recruited 200 participants from Amazon Mechanical Turk (MTurk) in return for payment. For the EEG component, 22 participants (15 females, 7 males; mean age 20, range: 18-28) were recruited from the University of Sydney in return for course credit. Subjects all had normal or corrected-to-normal vision and had no reported history of psychiatric or neurological disorders. The study was approved by the Human Ethics Committee of the University of Sydney. Verbal and written consent was obtained from each participant.

Stimuli & Design

GAN-generated stimuli were obtained from StyleGAN output found at shorturl.at/josOY (Karras et al., 2019). For a full description of the StyleGAN generative procedure and output, see Karras et al. (2019). Fake stimuli consisted of 25 faces, cars, and bedrooms at truncation levels of Ψ0.5 (realistic) and Ψ1.0 (unrealistic), (Figure 1B). To best match image statistics across real and fake images, real images were obtained from training images used for GAN output. These real training faces were obtained from the Flickr-Faces-HQ dataset (Karras et al., 2019). Real cars and bedrooms were randomly selected from the LSUN dataset (Yu et al., 2015). To maintain consistent aspect ratios, all images were cropped to a square aspect ratio and resized to a 256 × 256 pixel dimension. No other filtering or editing was applied to the stimuli to provide a naturalistic demonstration of visual processing. To reduce obvious surface-level inconsistencies between real and fake images, real faces with eyes not facing frontward and/or with overly pronounced facial expressions (e.g., crying, laughing) were excluded. Upon surface inspection, we found no consistent delineating features between the real and fake bedrooms and cars. All images were presented in both upright and inverted orientations totalling 450 stimuli overall (Figure 1B). To examine low-level properties at the image level, we took the mean image from the face, car, and bedroom stimuli and from 100 novel stimuli separately and computed the absolute pixel difference (Figure 1C). This allowed us to compare how visually distinct the face categories, i.e., how different the average unrealistic/realistic faces in our study were from an average real face.

Behavioural testing for real versus fake face discrimination was conducted online (for online/offline comparability see Grootswagers, 2020). The experiment was programmed in jsPsych (De Leeuw, 2015) and hosted on Pavlovia.org (Peirce, 2019). Two hundred participants performed real or fake face judgements for one of four comparisons (50 in each group): 1) upright unrealistic vs upright real, 2) upright realistic vs upright real, 3) inverted unrealistic vs inverted real, and 4) inverted realistic vs inverted real. Before beginning, participants were informed that half of the faces were fake, and half were real. They were not given any information on how to tell the difference. To investigate how untrained observers would identify real from fake faces given only a brief glimpse, we were interested in purely naïve observation. Each observer was shown 50 images in total: 25 fake and 25 real. Participants were informed that 50% of the images were real photos and 50% were computer-generated and were instructed to choose whether each image was real or fake. Each image was individually presented on the screen for 200ms, followed by a blank screen until the participant pressed a button to indicate if the face was real or fake. Stimuli were presented at 256 x 256 pixel dimension against a grey background. Presentation of images was randomised, and each image was only presented once. The experiment took around 3-5 minutes to complete (Figure 1D).

For the EEG component, the experiment was presented in Psychopy2 (Peirce et al., 2019). Participants sat in a dimly lit room approximately 60cm away from a 1920 x 1080 pixel Asus computer monitor. Stimuli subtended approximately 6.4 degrees visual angle on a grey background with a white fixation circle superimposing the stimuli at approximately 1.3 degrees. Images were presented at a 256 x 256 pixel dimension and in a rapid serial visual presentation (RSVP) paradigm, whereby stimuli are presented in rapid succession, at 20Hz and 5Hz sequences (33ms image duration and 167ms or 17ms gap). There were 20 sequences at each presentation rate comprising 40 in total with 18,000 images presented overall (with 20 repeats of each stimulus at each presentation rate). A sequence was started with a button press and lasted approximately 40 seconds. Subjects were instructed to fixate upon a white circle superimposed over each stimulus at the centre of the screen and told to respond by pressing any button on a 4-way button box whenever they spotted the fixation circle turn red (Figure 1D). Fixation colour changes were randomised to occur between 2 and 5 times in each sequence. Length of colour change corresponded to the time of one image presentation (33ms). At the conclusion of the experiment, participants were debriefed and informed that half the images had been fake.

EEG recordings and preprocessing

Continuous EEG data were recorded using a 64-electrode Brain Products EEG cap (Standard 64Ch actiCAP; GmbH, Herrsching, Germany) at a sample rate of 1000-Hz. Ag/AgCl active electrodes were placed in accordance with a 10/20 international system (Oostenveld & Praamstra, 2001). Electrode gel was applied to the scalp under each electrode, aiming to reduce signal impedances to below 10kΩ. Stimulus onset was synchronised to the EEG using transistor-transistor logic (TTL) pulses from the stimulus presentation computer to a separate recording computer. Pre-processing of the EEG data was computed offline using EEGLAB (Delorme & Makeig, 2004). The continuous EEG data were filtered with a high-pass filter of 0.1-Hz and a low-pass filter of 100-Hz and re-referenced to the average of all electrodes. No notch filter was applied. The data were then separated into epochs corresponding to stimulus presentation ranging from 100ms to 1000ms pre- and post-stimulus onset. This produced 180,000 pre-processed epochs for each participant.

Decoding analysis

Time-resolved MVPA decoding analysis of EEG data was implemented in MATLAB with the CoSMoMVPA toolbox (Oosterhof, Connolly, & Haxby, 2016). We used Linear Discriminant Analysis (LDA) classifiers as implemented in CoSMoMVPA in a leave-one-out cross-validation scheme. The LDA classifier estimated the probability of EEG data belonging to a certain group (e.g., real or fake) where the higher estimate is the predicted class (Grootswagers, Wardle, & Carlson, 2017). This was repeated at every time point, for every exemplar, and averaged across subjects to generate the mean cross-validation decoding performance at each time point. Classification performance was characterized as significant if it produced an above-chance accuracy (>50% for real versus fake decoding or 33% for 3-way category decoding). An above-chance decoding accuracy informs us that the EEG data contains information relevant the contrast of interest (Grootswagers, Wardle, & Carlson, 2017; Olivetti et al., 2012; Pereira et al., 2009).

Category Decoding Analysis

We performed a category decoding analysis to investigate whether there were meaningful differences among the face, car, and bedroom stimuli. We used an image-by-sequence cross-validation approach (Grootswagers, Robinson, & Carlson, 2019), which entailed training the classifier on all-but-one image from each of the three categories from all-but-one sequence and testing the classifier on left-out images from the left-out sequence. This ensured that the classifier had to generalize to novel exemplars to successfully decode between faces, cars, and bedrooms for each of the real, realistic, and unrealistic conditions (Carlson et al., 2013). Decoding accuracy was characterized by an above-chance classifier performance (>33%). Contrasts were broken down into presentation rate (5-Hz or 20-Hz), realism level (real, unrealistic, realistic), and configuration (upright, inverted).

Real versus Fake Decoding Analysis

We investigated whether real and fake image differences could be decoded from the EEG data using a leave-one-out cross validation approach. The leave-one-out cross-validation approach consists of dividing the data into training and testing sets whereby the classifiers are trained on all stimuli but one pair of real and fake stimuli from all but one RSVP sequence and then tested on the left-out stimulus pair from the remaining sequence. This ensured that the classifier had to generalise to the novel stimulus to successfully decode the category (i.e., real or fake) and could not rely on individual image-specific properties. Real stimuli were decoded against fake stimuli. Contrasts were broken down into presentation rate (5-Hz or 20-Hz), realism level (unrealistic, realistic), and configuration (upright, inverted). Thus, there were 8 decoded contrast combinations per image category. Given the large face processing literature and our clear hypotheses regarding faces, we were mainly interested in fake versus real decoding of faces; results from the car and bedroom categories are included for completeness in the Supplementary Materials and on https://osf.io/n2z73/.

To map the spatial distribution of the signal, we repeated the real versus fake decoding analysis at separate locations on the scalp. For each channel, we selected the four closest neighbouring channels and performed the exact same decoding analysis described above on just this local cluster of channels, storing the resulting accuracies at the centre channel. This results in a channel topography of decoding results that provides insight into the spatial origins of the signal.

As an exploratory follow-up analysis, we examined the relationship between real-fake decoding accuracy from the EEG data and behavioural accuracies obtained from the online participants for each individual image (Grootswagers et al., 2018; Ritchie et al., 2019). For each subject and each time point in the real-fake decoding analysis, we correlated (Spearman's rho) the image-specific average classifier accuracies with their corresponding behavioural accuracies. We then performed group level inference on the resulting subject-wise time-varying brain-behaviour correlations. If successful real/fake decoding in EEG reflects the real/fake signal that is 'used' by the brain to guide behaviour (Grootswagers et al., 2018; Ritchie et al., 2019), then we would expect a positive correlation between image-specific EEG-classification accuracy and behavioural accuracy. That is, faces identified as real or fake by the classifier would also be identified as real or fake by the participants.

Statistical inference

For the decoding and behavioural analyses, we used Bayesian statistics to characterize evidence arising from the data as either supporting the presence (alternative hypothesis) or absence (null hypothesis) of an effect. (Dienes, 2011; Jeffreys, 1998; Rouder et al., 2009; Wagenmakers, 2007). We used a standard JZS prior to calculate the null and alternative hypotheses (Rouder et al., 2018), which is a Cauchy distribution with a scale factor of 0.707 to determine the evidence of above-chance performance (e.g., >50% decoding) and a null-hypothesis point prior at chance-level (Morey & Rouder, 2011). For ease of interpretation, we thresholded Bayes factor (BF) values > 10 for strong evidence for the alternative hypothesis and BF values < ⅓ as evidence in favour of the null hypothesis (Morey & Rouder, 2011). For the decoding analyses, BFs serve as continuous degrees of evidence across multiple time points and not specific hypothesis testing at single time points. Thus, isolated BFs at single time points which did not reach threshold were not treated as evidence for either hypothesis if the surrounding points did not reach threshold or were interspersed with below-threshold values. Rather, BFs were treated as evidence if surrounding points were at threshold (Mai et al., 2019). For the decoding analyses, we, in addition, computed corresponding frequentist statistics using sign-permutation tests (1000 permutations) and Monte-Carlo cluster statistics with TFCE as cluster-statistic (Smith & Nichols, 2009), corrected for multiple comparisons across time using the max-statistic method (Maris & Oostenveld, 2007).

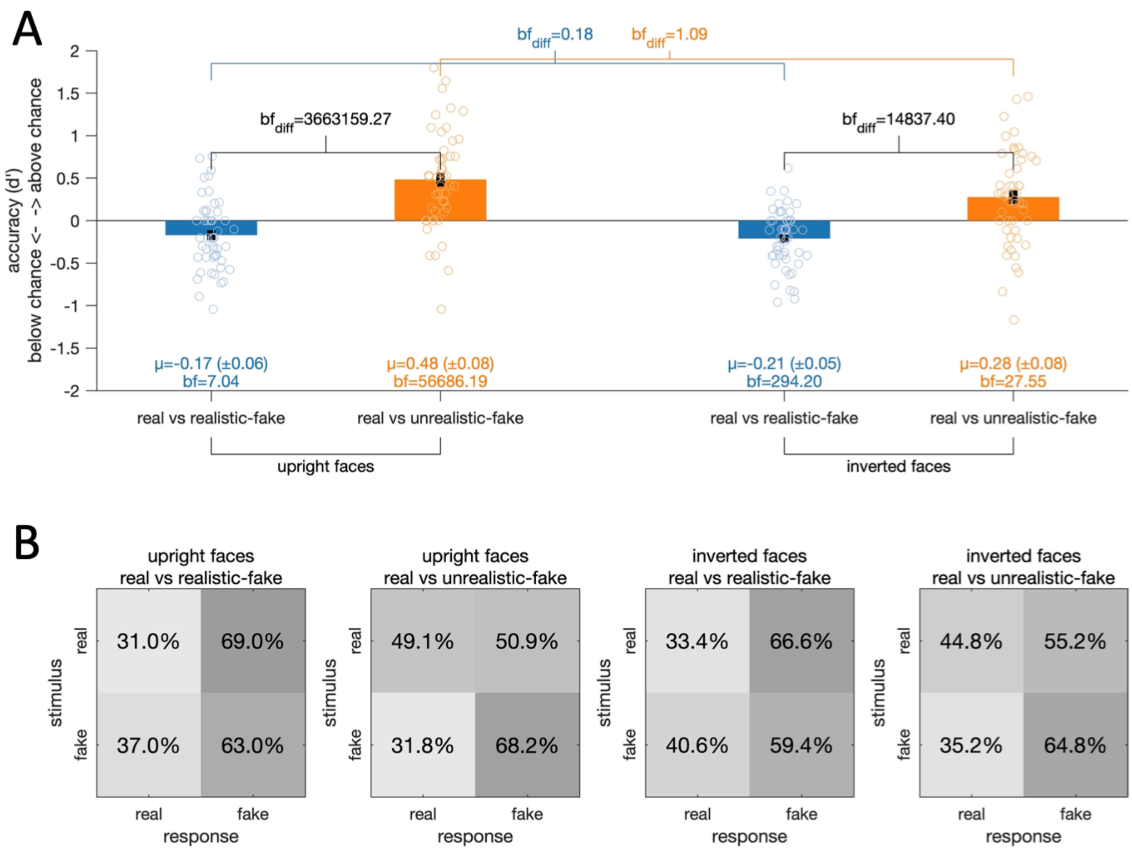

Figure 2. Behavioural discrimination of real and fake faces. A) In an upright (left) and inverted (right) configuration, discriminability for real/realistic (blue) faces was below chance, but above chance for real/unrealistic faces (orange). Performance was similar regardless of whether faces were upright or inverted.

Bars show mean and standard error. Each circle represents the response of one subject in one condition. The Bayes Factors (displayed above the x-axis) compute the evidence for a difference from chance discriminability (50% accuracy), and difference between conditions (stimulus and orientation). B) Confusion matrices display the results from the 4 behavioural categorisation conditions.

Results

Behavioural Performance

We were interested in whether participants could correctly classify and discriminate between real and fake faces. We calculated the proportion of images that were judged correctly as real or fake for each of the realistic/unrealistic and upright/inverted conditions and aggregated the judgements over participants. The main findings are presented in Figure 2. As indexed by a d' discriminability analysis, we found that participants could reliably discriminate real from unrealistic fake faces (0.48 ± 0.08, BF = 57.19) but could not discriminate real from realistic fake faces (-0.17 ±0.06, BF =7.04). Orientation had little effect on discriminability. Confusion matrices (Figure 2B) indicate that observers could correctly classify realistic fake faces (63%, se = 0.026, BF > 100) and unrealistic fake faces (68.2%, se = 0.026, BF > 100). However, observers performed at chance (50.9%, se = 0.027, BF = 0.16) when it came to classifying real faces from unrealistic fake faces and well below chance at classifying real faces from realistic fake faces (31%, se = 0.023, BF > 100). Classification performances were similar for inverted faces.

Overall, this demonstrates that although observers could reliably spot the fakes, they performed poorly at correctly labelling the real faces. Interestingly, participants had a below-chance discriminability for real and realistic fake faces. That is, observers overwhelmingly perceived realistic fake faces as appearing *more real* than the real faces consistent with other findings (Sanders et al., 2019). Importantly, inverting the faces had little effect on discriminability suggesting that detection was not reliant on configural or featural information (Tanaka et al., 2014).

Categorical decoding analysis

To examine whether real and fake images evoked similar categorical decoding effects compared to the previous literature, we decoded image category (cars, faces, and bedrooms) at all levels of realism (real, realistic, unrealistic), (Figure 3). As expected, we observed similar category-related dynamics for the real, realistic, and unrealistic images across all conditions. At a 5Hz presentation rate, we observed above-chance decoding for all categories at real, realistic, and unrealistic (Figure 3A). Decoding emerged and remained above-chance from 100ms until 700ms post-stimulus onset with an early peak at 120ms, a second peak at 200ms and a third peak at 250ms-300ms.

We then tested how category decoding was affected by our control manipulations (inversion and presentation rate). We observed similar above-chance decoding for all categorical and realism levels upon inversion (Figure 3B) and at a 20Hz presentation rate (Figure 3C), albeit less pronounced with simultaneous stimulus inversion and 20Hz presentation (Figure 3D). When upright and inverted, faces, cars, and bedrooms could be decoded at all levels of realism with similar temporal dynamics reported elsewhere (Grootswagers, Robinson, & Carlson, 2019; Grootswagers, Wardle, & Carlson, 2017).

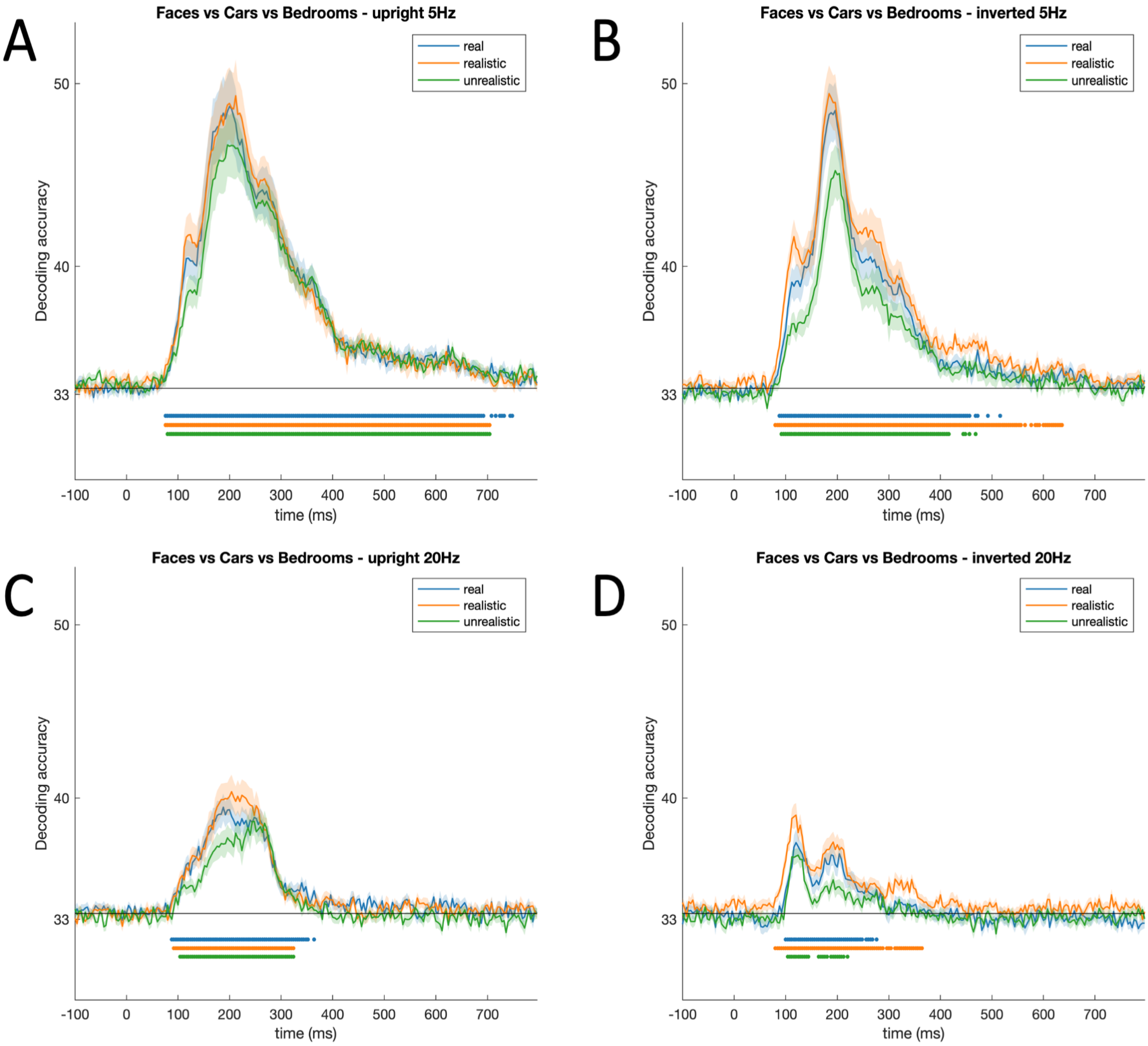

Figure 3. **Summary of category decoding using orientation and presentation rate manipulation.** A classifier was trained on EEG data from all categories, orientations, and presentation rates. Above-chance distinct category decoding was found for real (blue), realistic (orange), and unrealistic (green) stimuli regardless of orientation, presentation rate or stimuli type. Lines represent decoding accuracy over time with shaded areas displaying standard error across subjects (N = 22). Thresholded p-values below 0.05 are displayed under each pot.

Decoding Realness from EEG: Real vs Fake Faces

To determine if the brain could distinguish real from fake, we then investigated differences in neural patterns evoked from real and fake faces. At 5Hz and upright (Figure 4A), above-chance decoding emerged and peaked for unrealistic faces at around 100ms, 200ms, and 300ms (BF > 10) and fell below-chance at approximately 370ms (BF < ⅓). This decodability is reflective of early, rapid, low-level image perception followed by a later, higher-level, holistic decoding consistent with the temporal unfolding of face perception (Dobs et al., 2019; Balas & Koldewyn, 2013; Mühlberger et al., 2009). For realistic fake faces, decoding emerged at around 170ms and remained above-chance until approximately 240ms (BF > 10), suggesting a higher-level basis for discrimination of realistic and real faces. Although observers had trouble distinguishing real from fake faces and tended to overclassify faces as fake, the EEG data contained signal information relevant to this distinction which meaningfully differed between realistic fakes and unrealistic fakes, and this signal appeared to be constrained to a relatively short stage of processing.

If the information that we were decoding at 5Hz was reliant on image features rather than a face-processing effect, then we would predict that we could achieve a similar decoding result on inverted faces. However, at 5Hz and inverted (Figure 4B), only unrealistic fake faces were decodable from real faces. Above-chance decoding emerged at around 100ms (BF > 10), peaked at around 170ms, and was at chance again at approximately 250ms (BF < ⅓). In contrast, realistic faces remained at-chance and were not decodable from real faces (BF < ⅓). This suggests that inversion, known to disrupt configural processing of faces, is similarly disrupting a face-specific mechanism accounting for decoding differences between realistic and unrealistic faces (Jacques, d'Arripe, & Rossion, 2007; Rossion et al., 2000).

An alternative way to disrupt face-processing is to use faster presentation rates (Collins, Robinson, & Behrmann, 2018). At 20Hz and upright (Figure 4C), above-chance decoding emerged for unrealistic faces at around 100ms and was sustained until approximately 170ms (BF >10). Decodability for realistic faces emerged at 170ms and remained above chance until around 230ms (BF >10), showing very similar dynamics to the upright condition. Faster presentation rates have been shown to limit the extent and capacity for visual processing (Robinson, Grootswagers, & Carlson, 2019), but this result suggests short presentations can still yield information informative of real versus fake face distinctions, albeit with numerically lower and less sustained decoding accuracy.

Lastly, at 20Hz and inverted (Figure 4D), decoding performance was at chance for realistic and unrealistic fake faces (BF < ⅓). This suggests that inversion plus a faster presentation rate is enough for the EEG data to no longer contain any relevant information pertaining to real versus fake face distinctions. In other words, configural processing has been disrupted to an extent that activity patterns evoked from fake faces were not differentiable from activity evoked from real faces. As expected, real versus fake bedroom and car decoding was not so evident and can be found on https://osf.io/n2z73/.

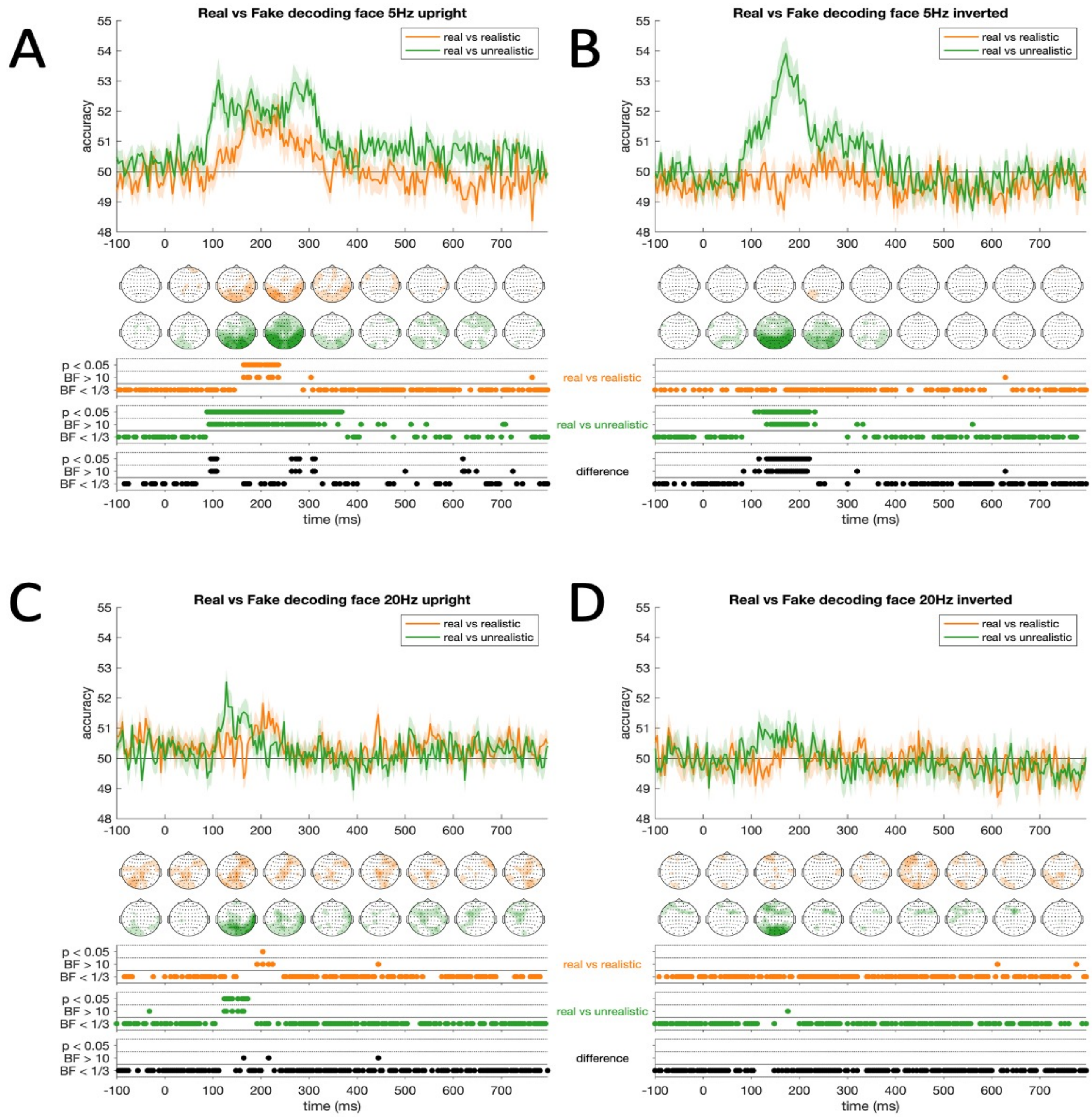

Figure 4. Decoding real versus fake faces. Different effects of orientation and presentation rate on decoding real and fake faces. Plots show decoding performance over time for real and fake (realistic or unrealistic) faces in upright and inverted orientations and at 5Hz and 20Hz presentation rates. The lines in each plot indicate classifier accuracy from time of stimulus onset until 800ms, with shaded areas showing standard errors across each subject (N = 22). Time-varying topographies are presented below for visualization purposes. Each plot was averaged across 100ms time bins where darker shades indicate contribution of channels to real/fake

decoding. In the lowest panel, thresholded p-values and Bayes Factors indicate above-chance decoding or non-zero differences.

Finally, we examined the relationship between real-fake decoding accuracy and behavioural classification accuracy across the images. If successful real/fake decoding in EEG reflects the real/fake signal that is 'used' by the brain to guide behaviour (Grootswagers et al., 2018; Ritchie et al., 2019) then we would predict to observe a positive correlation between image-specific EEG-classification accuracy and behavioural accuracy. Figure 5 shows the time-varying correlations for the upright and inverted 5Hz conditions. We did not perform this analysis for the 20Hz conditions due to limited above-chance decoding. We observed evidence for a positive brain-behaviour correlation around 170ms for the upright and inverted unrealistic faces, which is consistent with time points of above-chance decoding (Figure 4A). This result suggests that, at least for the unrealistic faces, the signal that is used by the classifier for real/fake distinction could be used by the brain to make the real/fake decision (Grootswagers, Cichy, & Carlson., 2018; Ritchie, Kaplan, & Klein, 2019).

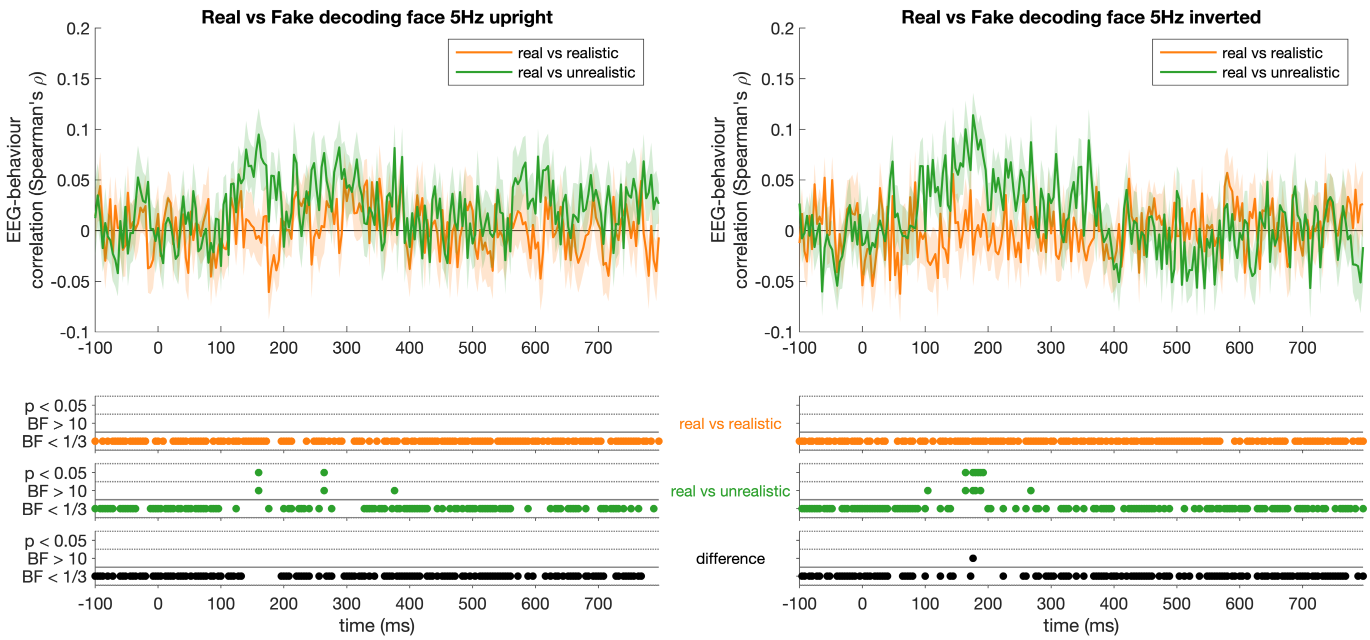

Figure 5. Correlating behavioural accuracy with decoding. Plots show the relationship between image-specific EEG decoding accuracy and behavioural accuracy over time for the 5Hz upright condition (left) and 5Hz inverted condition (right). The lines indicate correlation from time of stimulus onset until 800ms for realistic versus real faces (orange) and unrealistic versus real faces (green), with shaded areas showing standard errors. In the lowest panel, thresholded p-values and Bayes Factors indicate above-chance correlation or non-zero differences. Positive brain-behaviour correlations can be seen at around 170ms and 270ms for upright unrealistic faces (green) and at around 150ms-200ms for inverted unrealistic faces (BF>10).

Discussion

There is growing concern that realism is advancing at such a rate that humans will have difficulty discerning between what is real and what is fake (Fletcher, 2018; Khodabakhsh et al., 2019; Nightingale et al., 2017; Shen et al., 2019). Our results demonstrate that given only a brief glimpse, observers may be able to spot fake faces. However, they have a harder time discerning real faces from fake faces and, in some instances, believed fake faces to be more real than real faces. However, using time-resolved EEG and multivariate pattern classification methods, we found that it was possible to decode both unrealistic *and* realistic fake faces from real faces using brain activity. This dissociation between behaviour and neural responses for realistic faces yields important new evidence about fake face perception as well as implications involving the increasingly realistic class of GAN-generated faces.

Our behavioural results are consistent with previous research that suggests that observers typically display difficulties with correctly discriminating between real and realistic fake faces despite face expertise (Holmes et al., 2016; Nightingale et al., 2017; Sanders et al., 2019; Zhou et al., 2019). For example, in a two-alternative forced-choice task, participants would judge realistic face masks as being more realistic than human faces on a third of all trials (Sanders et al., 2019). Artificial faces made by GANs have also recently received attention and have been similarly demonstrated to fool observers (Hulzebosch et al., 2020; Isola et al., 2017; Zhou et al., 2019; Liu et al., 2020). As expected, we found that it was harder to spot the realistic fake faces than the unrealistic fake faces, although observers were able to correctly classify the fakes. However, participants struggled at discriminating real from realistic fakes and overclassified fake faces as being real. We presented faces for 200ms, which could be considered a brief exposure period, but the images were not masked so processing would have continued even after the images had disappeared (Robinson, Grootswagers, & Carlson, 2019). Given a long enough time to observe, Liu et al., (2020) found that identifying artifacts such as "asymmetrical eyes" and "irregular teeth" in artificial faces can assist in spotting fakes. Presumably, assessing such details requires more time and eye movements. Indeed, observers can be trained to reliably spot fake faces by learning what to look for (Hills & Lewis, 2006; Tanaka & Farah, 1993). Here, our primary focus was examining first impression responses of naïve observers by limiting the time spent looking at each face and giving participants unlimited time to make a response. With some added training it remains to be seen whether observers may be able to use that information to make a more accurate decision. Future studies may investigate whether training observers on GAN-generated faces and whether allowing for longer stimulus durations enhances detection.

We found that although observers may have difficulties discriminating between real and realistic faces, they have distinct representations in the human visual system. Given that category decoding was most pronounced and sustained in the 5Hz and upright condition, enough for each image to reach a high-level representation in the brain (Grootswagers, Robinson, & Carlson, 2019), we expected real/fake decoding to be most pronounced in this condition too. Above-chance decoding represents the classifier successfully distinguishing neural activity evoked from real and fake faces, namely, real/fake differences. Critically, a leave-one-out cross validation approach (see methods) ensured that the classifier could not learn to categorise the EEG data based on visual features or low-level properties belonging to specific faces, but rather had to generalize learned category information (real/fake) onto novel stimuli (Carlson et al., 2013; Grootswagers, Wardle, & Carlson, 2016; Teichmann et al., 2020). This guaranteed that the classifier performance related to a group-level distinction rather than to individual image-level properties.

Indeed, for the 5Hz, upright condition, we found that the classifier successfully discriminated between unrealistic/real as well as realistic/real faces (Figure 4A). Decoding for unrealistic faces displayed a triple peak pattern, emerging at around 100ms maintained until around 370ms. Early decoding differences are consistent with rapid face detection and face-specific processing (Rossion et al., 2015; Dobs et al., 2019; Crouzet, Kirchner, & Thorpe, 2010; Wardle et al., 2020). The latter two peaks (at around 170-200ms and 270-320ms) have been similarly demonstrated to emerge in real versus artificial face perception (Wheatley et al., 2011; Balas & Koldewyn, 2013; Sagiv & Bentin, 2001; Schindler et al., 2017, Schindler et al., 2019, Wardle et al., 2020). Schindler et al (2017) suggest that early-stage N170 processing is related to assessing the structural configuration of faces as seen by a greater occipital involvement whilst the later-staged LPP, seen to increase linearly with face realism, suggests a deeper person-related, semantic involvement (also see Abdel Rahman, 2011, Taylor, Shehzad, & McCarthy, 2016). Differences at the triple peak correspond to N250 and P300 components typically associated with face familiarity (Collins et al., 2018) and semantic information (Tanaka et al., 2006), the latter especially important for behaviour (Hanso et al., 2010). In contrast, realistic/real decoding displayed a single-peak emergence between around 170ms to 240ms indicating a difference in processing between realistic and unrealistic faces. Namely, that differences in perception between real and realistic faces were constrained to the 170ms time period. Indeed, in comparing human faces to doll faces and artificial faces, others have shown that only the human faces typically evoke sustained neural responses beyond the N170 component necessary for higher-order perception (Balas & Koldewyn, 2013; Wheatley et a., 2011). Balas and Koldewyn (2013) found that the N170 was better characterised by encoding deviations from facial appearance than it was for animacy perception. In other words, realistic faces were perceived as configurally different to real faces, but that only unrealistic faces engaged later processing necessary for high-order animacy or familiarity perception. Overall, earlier decoding for unrealistic faces, consistent with apparent low-level image differences (Figure 1C), suggests that early and low to mid-level processing differences may account for decodability between real and unrealistic faces. The decoding for realistic faces, by contrast, emerges later and is constrained to the 170ms time period, suggesting a face-specific configural process may be responsible for this distinction.

Assessing fake/real decoding for inverted faces allows us to evaluate whether the fake/real distinction relies on mechanisms that are responsible for the superiority in face recognition for upright faces relative to inverted faces. Inversion disrupts the configural processing of faces by making them appear more like objects whilst retaining low-level stimulus attributes (Eimer, 2000; Leder & Bruce, 2000; Rousselet et al., 2003). Firstly, we found that inversion led to the disruption of decoding for realistic faces (Figure 4B). In contrast, we found that decoding for unrealistic inverted faces was preserved but less sustained when compared to upright. The peak in decoding may be reflective of increased featural processing for inverted unrealistic faces, also seen to occur with distorted or 'Thatcherized' faces (Carbon et al., 2005; Milivojevic et al., 2003). Lack of above-chance decoding for inverted realistic faces may reflect the contribution of high-level, expertise-driven capabilities for upright fake face detection when face processing mechanisms, rather than object processing, were available. Overall, we found that upon stimulus inversion our decoding results were consistent with a face-specific or expertise response, such that realistic fake faces could not be discriminated from real faces when typical face perception was disrupted, even though the same visual features were present.

The presentation of images at a faster presentation rate limits the consolidation of each image and build-up of higher-order representation (Grootswagers, Robinson, & Carlson, 2019)., allowing an analysis of the contribution of low-level processing. At a faster presentation rate of 20Hz, we found that upright fake faces could be discriminated from real faces for the realistic and unrealistic conditions (Figure 4C). Indeed, early, low-level visual processing is fairly unaffected by image presentation durations (Grootswagers, Robinson, & Carlson, 2019). Observing less sustained decoding is consistent with the limited capacity and extent of visual processing since each image is masked by every successive image to a greater extent and therefore places limits on visual processing compared to a slower presentation rate (Collins, Robinson, & Behrmann, 2018; Robinson, Grootswagers, & Carlson, 2019). Additionally, higher-level, identity or semantically related face information discernible in the slow condition was possibly limited at the faster presentation rate consistent with Collins et al. (2018). In sum, we found that unrealistic faces could be decoded upon inversion and at a faster presentation rate suggesting the contribution of low-level visual differences. By contrast, we could not decode realistic faces when inverted, but we could decode at a faster presentation rate, indicating that fake/real perception was likely driven by expertise and face-specific processing.

Interestingly, we found that neural differences between real and realistic fake faces did not translate into a reliable behavioural decision for realistic face discrimination at the population level. We found a brain-behaviour correlation at around 150ms-200ms for unrealistic versus real faces, suggesting that this time period of processing is important for behaviour. However, the same correlation was not observed for the realistic faces. One possibility is that whilst our data indicates that a realistic fake/real signal is present, this signal gets 'lost' in the visual hierarchy and consequently remains uninformative for behaviour. For instance, although animacy categorisation can be decoded throughout the entire ventral visual stream, this information is most suitably formatted for behaviour in higher-level visual areas like the ventral occipital and parahippocampal cortex (Grootswagers, Cichy, & Carlson, 2018). Since decoding unrealistic/real faces was more sustained than realistic/real faces, associated more with in-depth face processing at later stages (i.e., LPP), it is possible that this level of extended processing is required for behavioural "readout" (see de-Wit et al., 2016; Grootswagers, Cichy, & Carlson, 2018; Ritchie, Kaplan, & Klein 2019). Yet, the highest brain-behaviour correlation for unrealistic faces was observed at 150-200ms, a time when decoding was not reliably different between the realistic and unrealistic condition. This has several implications. In an applied setting such as cyber security or Deepfakes, examining the detection ability for realistic fake faces might be best pursued using machine learning classifiers applied to neuroimaging data rather than targeting behavioural performance. As we have shown, the former contains discriminative relevance whereas observers may actually perform worse than chance given the decision (and a brief glance). A third related possibility is that the decodable real/fake face signal is operating below conscious access and therefore is not picked up by our behavioural task. This is reminiscent of findings that individuals with prosopagnosia who cannot behaviourally classify or recognise faces as familiar or unfamiliar nevertheless display stronger autonomic responses to familiar faces than unfamiliar faces (Tranel & Damasio, 1985). Similarly, what we have shown in this study is that whilst we could accurately decode the difference between real and realistic faces from neural activity, that difference was not seen behaviourally. Instead, observers incorrectly identified 69% of the real faces as being fake. Still, it is possible that a different behavioural task may have yielded a better performance. Forced to respond via a two-alternative forced-choice task or an implicit task such as face familiarity or trustworthiness may have engaged different behavioural processes more conducive for real/fake face discrimination. For instance, behaviourally categorising faces as threatening, competent, or trustworthy has been shown to occur as quickly as 33- 100ms after onset (Bar et al., 2006; Willis & Todorov, 2006). Conversely, real or fake judgments may occur as late as 240ms after stimulus presentation (Zhou et al., 2019). Therefore, future work could investigate whether judgments about face trustworthiness or threat may be a better cue for detection than real or fake.

In sum, we found that there is a dissociation between the ability of participants to categorise faces as real or fake and the decodability of this distinction in the brain. In other words, although the brain can 'recognise' the difference between real and realistic fake faces, observers cannot consciously tell them apart. Our findings of the dissociation between brain response and behaviour have implications for the ways in which we study fake face perception, the questions we pose when asking about fake image identification, and the possible ways in which we can establish protective standards against fake image misuse.

Future studies may investigate the contribution of face expertise for decoding and behaviour. Expertise influences how deeply and configurally a face is perceived allowing for more subtle identification of spatial relations, features, and same-race faces (Wong et al., 2009; Tanaka, 2001; Tanaka & Taylor, 1991; Hancock & Rhodes, 2008; Meissner & Brigham, 2001). Indeed, individuals with digital manipulation training and experience (i.e., photo-editing and photography) are more able to identify fake images than non-experienced individuals (Shen et al., 2019). Having the same participants participate in both the EEG and behaviour experiments may be useful in exploring inter-individual differences and the influence of expertise.

In conclusion, we investigated to what extent state-of-the-art GAN faces made by AI fool human observers. Using behavioural and neuroimaging methods we found that it was possible to reliably detect AI-generated fake images using EEG activity given only a brief glance, even though observers could not consciously report seeing differences. Given that observers are already struggling with differentiating between fake and real faces, it is of immediate and practical concern to further investigate the important ways in which the brain is able to tell the two apart. It is becoming increasingly possible to rapidly and effortlessly generate realistic fake images, videos, writing, and multimedia that are practically indiscernible from real (Radford et al., 2019; Maras & Alexandrou, 2018; Asensio et al., 2014; Ledig et al., 2017). This capacity is only going to become more widespread and has profound implications for cybersecurity, fake news, detection bypass, and social media (Damiani, 2019; Fletcher, 2018; Maddocks, 2020). Already, a newer and more realistic set of images and faces have been generated by GANs that might challenge human perception more drastically than we have investigated here (Karras et al., 2020). Understanding the dissociation between brain and behaviour for fake face detection will have practical implications for the way we tackle the potentially detrimental and universal spread of artificially generated information.

Acknowledgements

The authors would like to thank Sophia Shatek for her help with EEG recording and Roy Raskolnikov for his suggestions and support. This research was supported by ARC DP160101300 (TAC), ARC DP200101787 (TAC), and ARC DE200101159 (AKR). The authors acknowledge the University of Sydney HPC service for providing High Performance Computing resources. The authors declare no competing financial interests.

References

Abdel Rahman, R. (2011). Facing good and evil: Early brain signatures of affective biographical knowledge in face recognition. *Emotion, 11*(6), 1397.

Adelani, D. I., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2020, April). Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *International Conference on Advanced Information Networking and Applications* (pp. 1341-1354). Springer, Cham.

Asensio, J. M. L., Peralta, J., Arrabales, R., Bedia, M. G., Cortez, P., & Peña, A. L. (2014). Artificial intelligence approaches for the generation and assessment of believable human-like behaviour in virtual characters. *Expert Systems with Applications, 41*(16), 7281-7290.

Axelrod, V., & Yovel, G. (2015). Successful decoding of famous faces in the fusiform face area. *PloS one, 10*(2), e0117126.

Balas, B., & Koldewyn, K. (2013). Early visual ERP sensitivity to the species and animacy of faces. *Neuropsychologia, 51*(13), 2876–2881. https://doi.org/10.1016/j.neuropsychologia.2013.09.014

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*(2), 269.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of cognitive neuroscience, 8*(6), 551-565.

Carbon, C. C., Schweinberger, S. R., Kaufmann, J. M., & Leder, H. (2005). The Thatcher illusion seen by the brain: an event-related brain potentials study. *Cognitive Brain Research, 24*(3), 544-555.

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. *Journal of vision, 13*(10), 1-1.

Carlson, T., Goddard, E., Kaplan, D. M., Klein, C., & Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage, 180*, 88–100. https://doi.org/10.1016/j.neuroimage.2017.08.019

Collins, E., Robinson, A. K., & Behrmann, M. (2018). Distinct neural processes for the perception of familiar versus unfamiliar faces along the visual hierarchy revealed by EEG. *NeuroImage, 181*, 120-131.

Crookes, K., Ewing, L., Gildenhuys, J. D., Kloth, N., Hayward, W. G., Oxner, M., ... & Rhodes, G. (2015). How well do computer-generated faces tap face expertise? *PloS one, 10*(11), e0141353.

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: face detection in just 100 ms. *Journal of vision, 10*(4), 16-16.

Damiani, J. (2019). A voice deepfake was used to scam a CEO out of $243,000.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods, 47(*1), 1-12.

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review, 23(*5), 1415–1428.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274-290.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods, 134*(1), 9-21.

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature communications, 10*(1), 1-10.

Eimer, M. (2000). Effects of face inversion on the structural encoding and recognition of faces: Evidence from event-related brain potentials. *Cognitive Brain Research, 10*(1-2), 145-158.

Farid, H., & Bravo, M. (2007). Photorealistic rendering: How realistic is it? *Journal of Vision, 7*(9), 766-766.

Farid, H., & Bravo, M. J. (2012). Perceptual discrimination of computer generated and photographic faces. *Digital Investigation, 8*(3-4), 226-235.

Farokhmanesh, M. (2018). Deepfakes Are Disappearing from Parts of the Web, But They're Not Going Away. *The Verge.*

Fletcher, J. (2018). Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance. *Theatre Journal, 70*(4), 455-471.

Gauthier, I., & Tarr, M. J. (2002). Unravelling mechanisms for expert object recognition: bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance, 28(*2), 431.

Gleicher, N. (2019). Removing Coordinated Inauthentic Behavior from Georgia, Vietnam, and the US. Facebook. *Retrieved from* http://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods, 52*, 2283–2286. https://doi.org/10.3758/s13428-020-01395-3

Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage, 179*, 252–262. https://doi.org/10.1016/j.neuroimage.2018.06.022

Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019). The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage, 188*, 668–679. https://doi.org/10.1016/j.neuroimage.2018.12.046

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2016). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience, 29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of cognitive neuroscience, 29(*4), 677-697.

Hancock, K. J., & Rhodes, G. (2008). Contact, configural coding and the other-race effect in face recognition. *British Journal of Psychology, 99*(1), 45-56.

Hanso, L., Bachmann, T., & Murd, C. (2010). Tolerance of the ERP signatures of unfamiliar versus familiar face perception to spatial quantization of facial images. *Psychology, 1*(03), 199.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. Nature Reviews *Neuroscience, 7*(7), 523-534.

Hills, P. J., & Lewis, M. B. (2006). Short article: reducing the own-race bias in face recognition by shifting attention. *Quarterly Journal of Experimental Psychology, 59*(6), 996-1002.

Holmes, O., Banks, M. S., & Farid, H. (2016). Assessing and improving the identification of computer-generated portraits. *ACM Transactions on Applied Perception (TAP), 13*(2), 1-12.

Hulzebosch, N., Ibrahimi, S., & Worring, M. (2020). Detecting CNN-Generated Facial Images in Real-World Scenarios. 642–643. https://openaccess.thecvf.com/content_CVPRW_2020/html/w39/Hulzebosch_Detecting_CNN-Generated_Facial_Images_in_Real-World_Scenarios_CVPRW_2020_paper.html

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-To-Image Translation with Conditional Adversarial Networks. 1125–1134. https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html

Jacques, C., d'Arripe, O., & Rossion, B. (2007). The time course of the inversion effect during individual face discrimination. *Journal of Vision, 7*(8), 3-3.

Jeffreys, H. (1998). The theory of probability. *OUP Oxford.*

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (*pp. 4401-4410).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8110-8119).

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135-146.

Khodabakhsh, A., Ramachandra, R., & Busch, C. (2019). Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX*) (pp. 1-6). IEEE.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *The quarterly journal of experimental psychology Section A, 53*(2), 513-536.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681-4690).

Liu, Z., Qi, X., & Torr, P. H. (2020). Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8060-8069).

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

Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 1-9.

Mader, B., Banks, M. S., & Farid, H. (2017). Identifying computer-generated portraits: The importance of training and incentives. *Perception, 46*(9), 1062-1076.

Mai, A. T., Grootswagers, T., & Carlson, T. A. (2019). In search of consciousness: Examining the temporal dynamics of conscious visual perception using MEG time-series data. *Neuropsychologia, 129*, 310-317.

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof, 23*(3), 255-262.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods, 164*(1), 177-190.

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018, April). Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 384-389). IEEE.

Milivojevic, B., Clapp, W. C., Johnson, B. W., & Corballis, M. C. (2003). Turn that frown upside down: ERP effects of thatcherization of misorientated faces. *Psychophysiology, 40*(6), 967-978.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological methods, 16*(4), 406.

Mühlberger, A., Wieser, M. J., Herrmann, M. J., Weyers, P., Tröger, C., & Pauli, P. (2009). Early cortical processing of natural and artificial emotional faces differs between lower and higher socially anxious persons. *Journal of neural transmission, 116*(6), 735-746.

Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., & Morishima, S. (2019). Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4480-4490).

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes? *Cognitive research: principles and implications, 2*(1), 30.

Olivetti, E., Veeramachaneni, S., & Nowakowska, E. (2012). Bayesian hypothesis testing for pattern discrimination in brain decoding. *Pattern Recognition, 45*(6), 2075-2084.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical neurophysiology, 112*(4), 713-719.

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics, 10*. https://doi.org/10.3389/fninf.2016.00027

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods, 51*(1), 195-203.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage, 45*(1), S199-S209.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *British Journal for The Philosophy of Science, 70*(2), 581–607. https://doi.org/10.1093/bjps/axx023

Robinson, A. K., Grootswagers, T., & Carlson, T. A. (2019). The influence of image masking on object representations during rapid serial visual presentation. *NeuroImage, 197*, 224–231. https://doi.org/10.1016/j.neuroimage.2019.04.050

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport, 11*(1), 69-72.

Rossion, B., Torfs, K., Jacques, C., & Liu-Shuang, J. (2015). Fast periodic presentation of natural images reveals a robust face-selective electrophysiological response in the human brain. *Journal of vision, 15(*1), 18-18.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review, 16(*2), 225-237.

Rousselet, G. A., Macé, M. J. M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of vision, 3*(6), 5-5.

Sagiv, N., & Bentin, S. (2001). Structural encoding of human and schematic faces: holistic and part-based processes. *Journal of cognitive neuroscience, 13*(7), 937-951.

Sanders, J. G., Ueda, Y., Yoshikawa, S., & Jenkins, R. (2019). More human than human: a Turing test for photographed faces. *Cognitive research: principles and implications, 4*(1), 1-10.

Schindler, S., Bruchmann, M., Bublatzky, F., & Straube, T. (2019). Modulation of face-and emotion-selective ERPs by the three most common types of face image manipulations. *Social cognitive and affective neuroscience, 14*(5), 493-503.

Schindler, S., Zell, E., Botsch, M., & Kissler, J. (2017). Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Scientific Reports, 7*(1), 45003. https://doi.org/10.1038/srep45003

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. Proceedings of the National *Academy of Sciences, 112*(41), 12887-12892.

Shen, C., Kasra, M., Pan, W., Bassett, G. A., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New media & society, 21*(2), 438-463.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter, 19(*1), 22-36.

Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2006). Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE, 94*(11), 1948-1962.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage, 44*(1), 83-98.

Sunday, M. A., Dodd, M. D., Tomarken, A. J., & Gauthier, I. (2019). How faces (and cars) may become special. *Vision research, 157*, 202-212.

Tanaka, J. W. (2001). The entry point of face recognition: evidence for face expertise. *Journal of Experimental Psychology: General, 130*(3), 534.

Tanaka, J. W., Curran, T., Porterfield, A. L., & Collins, D. (2006). Activation of preexisting and acquired face representations: the N250 event-related potential as an index of face familiarity. *Journal of Cognitive Neuroscience, 18*(9), 1488-1497.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly journal of experimental psychology, 46*(2), 225-245.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology, 23*(3), 457-482.

Tanaka, J. W., Kaiser, M. D., Hagen, S., & Pierce, L. J. (2014). Losing face: impaired discrimination of featural and configural information in the mouth region of an inverted face. Attention, *Perception, & Psychophysics, 76*(4), 1000-1014.

Taylor, J., Shehzad, Z., & McCarthy, G. (2016). Electrophysiological correlates of face-evoked person knowledge. *Biological psychology, 118*, 136-146.

Teichmann, L., Quek, G. L., Robinson, A. K., Grootswagers, T., Carlson, T. A., & Rich, A. N. (2020). The influence of object-color knowledge on emerging object representations in the brain. *Journal of Neuroscience, 40*(35), 6779-6789.

Tranel, D., & Damasio, A. R. (1985). Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. *Science, 228*(4706), 1453-1454.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review, 14(*5), 779-804.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8798-8807).

Wardle, S. G., Taubert, J., Teichmann, L., & Baker, C. I. (2020). Rapid and dynamic processing of face pareidolia in the human brain. *Nature Communications, 11*(1), 4518. https://doi.org/10.1038/s41467-020-18325-8

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science, 17(*7), 592-598.

Wheatley, T., Weinberg, A., Looser, C., Moran, T., & Hajcak, G. (2011). Mind Perception: Real but Not Artificial Faces Sustain Neural Activity beyond the N170/VPP. *PLOS ONE, 6*(3), e17960. https://doi.org/10.1371/journal.pone.0017960

Wong, A. C. N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects: Becoming a Ziggerin expert—but which type? *Psychological Science, 20*(9), 1108-1117.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint* arXiv:1506.03365.

Yu, J., Xu, X., Gao, F., Shi, S., Wang, M., Tao, D., & Huang, Q. (2020). Toward Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs. *IEEE Transactions on Cybernetics*.

Zhou, S., Gordon, M., Krishna, R., Narcomey, A., Fei-Fei, L. F., & Bernstein, M. (2019). Hype: A benchmark for human eye perceptual evaluation of generative models. In *Advances in Neural Information Processing Systems* (pp. 3449-3461).