

Our Subconscious Deepfake-Detection Skills Could Power Future Automated Systems

Originally published at <https://www.unite.ai/our-subconscious-deepfake-detection-skills-could-power-future-automated-systems/>

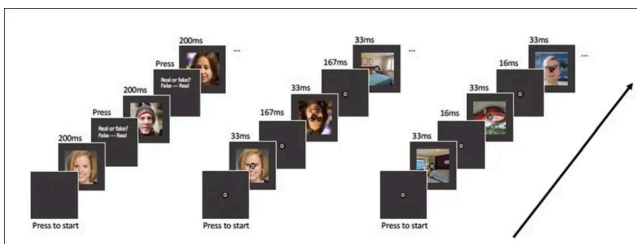


New research from Australia suggests that our brain is adroit at recognizing sophisticated deepfakes, even when we believe consciously that the images we're seeing are real.

The finding further implies the possibility of using people's neural responses to deepfake faces (rather than their stated opinions) to train automated deepfake detection systems. Such systems would be trained on images' deepfake characteristics not from confused estimates of plausibility, but from our instinctive perceptual mechanisms for facial identity recognition.

'[A]lthough the brain can 'recognise' the difference between real and realistic faces, observers cannot consciously tell them apart. Our findings of the dissociation between brain response and behaviour have implications for how 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.'

The results emerged in rounds of testing designed to evaluate the way that people respond to false imagery, including imagery of manifestly fake faces, cars, interior spaces, and inverted (i.e. upside down) faces.



Various iterations and approaches for the experiments, which involved two groups of test subjects needing to classify a briefly-shown image as 'fake' or 'real'. The first round took place on Amazon Mechanical Turk, with 200 volunteers, while the second round involved a smaller number of volunteers responding to the tests while hooked up to EEG machines. Source: https://tjil.github.io/tjil-grootswagers-pdf/Moshel_et_al_-_2022_-_Are_you_for_real_Decoding_realistic_AI-generated_.pdf

The paper asserts:

'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 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.'

The paper suggests that the new work has 'several implications' in applied cybersecurity, and that the development of deepfake learning classifiers should perhaps be driven by subconscious response, as measured on EEG readings in response to fake images, rather than by the viewer's conscious estimation of the veracity of an image.

The authors comment*:

'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](#).

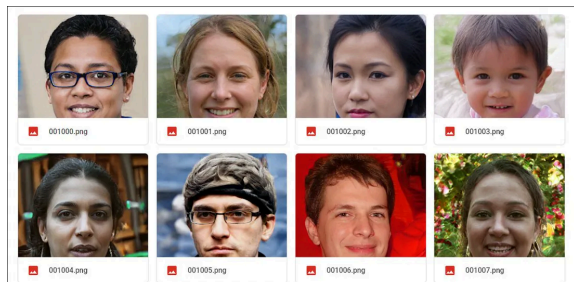
'Similarly, what we have shown in this study is that whilst we could accurately decode the difference between real and realistic faces from neural activity, that difference was not seen behaviourally. Instead, observers incorrectly identified 69% of the real faces as being fake.'

The [new work](#) is titled *Are you for real? Decoding realistic AI-generated faces from neural activity*, and comes from four researchers across the University of Sydney, Macquarie University, Western Sydney University, and The University of Queensland.

Data

The results emerged from a broader examination of human ability to distinguish manifestly false, hyper-realistic (but still false), and real images, carried out across two rounds of testing.

The researchers used images created by Generative Adversarial Networks (GANs), [shared](#) by NVIDIA.

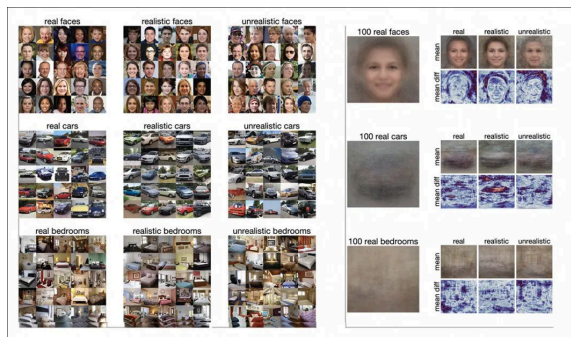


GAN-generated human face images made available by NVIDIA. Source: <https://drive.google.com/drive/folders/1EDYEYR3IB71-5BbTARQkhg73leVB9tam>

The data comprised 25 faces, cars and bedrooms, at levels of rendering ranging from 'unrealistic' to 'realistic'. For face comparison (i.e. for suitable non-fake material), the authors used selections from the source data of NVIDIA's source Flickr-Faces-HQ (FFHQ) [dataset](#). For comparison of the other scenarios, they used material from the [LSUN](#) dataset.

Images would ultimately be presented to the test subject either the right way up, or inverted, and at a range of frequencies, with all images resized to 256x256 pixels.

After all material was assembled, 450 stimuli images were curated for the tests.



Representative examples of the test data.

Tests

The tests themselves were initially carried out online, through [jsPsych](#) on pavlovia.org, with 200 participants judging various subsets of the total gathered testing data. Images were presented for 200ms, followed by a blank screen that would persist until the viewer made a decision as to whether the flashed image was real or fake. Each image was only presented once, and the entire test took 3-5 minutes to complete.

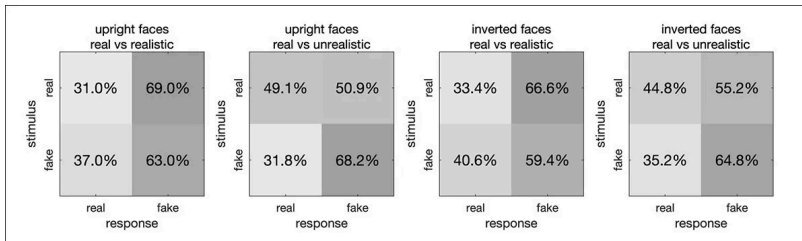
The second and more revealing round used in-person subjects rigged up with EEG monitors, and was presented on the [Psychopy2](#) platform. Each of the twenty sequences contained 40 images, with 18,000 images presented across the entire tranche of the test data.

The gathered EEG data was decoded via MATLAB with the CoSMoMVP toolbox, using a [leave-one-out cross-validation](#) scheme under Linear Discriminant Analysis ([LDA](#)).

The LDA classifier was the component that was able to make the distinction between the brain reaction to fake stimuli, and the subject's own opinion on whether the image was fake.

Results

Interested to see whether the EEG test subjects could discriminate between the fake and real faces, the researchers aggregated and processed the results, finding that the participants could discern real from unrealistic faces easily, but apparently struggled to identify realistic, GAN-generated fake faces. Whether or not the image was upside down appeared to make little difference.

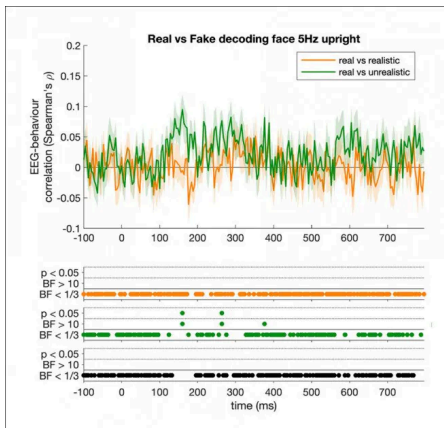


Behavioral discrimination of real and synthetically-generated faces, in the second round.

However, the EEG data told a different story.

The paper states:

'Although observers had trouble distinguishing real from fake faces and tended to overclassify fake faces, the EEG data contained signal information relevant to this distinction which meaningfully differed between realistic and unrealistic, and this signal appeared to be constrained to a relatively short stage of processing.'



Here the disparity between EEG accuracy and the reported opinion of the subjects (i.e. as to whether or not the face images were fake) are not identical, with the EEG captures getting nearer to the truth than the manifest perception of the people involved.

The researchers conclude that although observers may have trouble tacitly identifying fake faces, these faces have 'distinct representations in the human visual system'.

The disparity found has caused the researchers to speculate on the potential applicability of their findings for future security mechanisms:

'In an applied setting such as cyber security or Deepfakes, examining the detection ability for realistic faces might be best pursued using machine learning classifiers applied to neuroimaging data rather than targeting behavioural performance.'

They conclude:

'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.'

* My conversion of inline citations to hyperlinks.

First published 11th July 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More