[dstl]
The Science Inside

The Alan Turing Institute

Lancaster University

[ ] Work with us
Please email us if you wish to be added to the mailing list.

[ ] Contribute
Have something to share, please let us know and we can feature it in a future issue.

[ ] Spread the Word
Please feel free to pass this onto your colleagues.

[ ] Contact Details
caiss@lancaster.ac.uk
caiss@dstl.gov.uk

To email us at Dstl, scan QR code

DSTL/TR150673

Issue: 5 **Date: July 2023**

CAISS
Computation & AI for Social Science Hub

Newsletter in collaboration with The Alan Turing Institute and Lancaster University

Deep fakes – a cause for concern?

In this issue we wanted to take a look at deep fakes and how easy it is to detect them. Image manipulation/editing is nothing new, and deep fakes are the latest in a long line of techniques used for manipulation. Joseph Stalin had people removed from photographic images of him so he was not seen to be associating with the "wrong type of people".

What is a Deep fake? Deep fakes refer to audio, image, text or video that have been automatically synthesised by a machine learning system and AI. Deep fake technology can be used to create highly realistic images or videos that may depict people saying or doing something that they did not. For example, recent images have circulated of the Pope wearing a large white "puffer" coat, something he never did. Link here: https://www.bloomberg.com/news/newsletters/2023-04-06/pope-francis-white-puffer-coat-ai-image-sparks-deep-fake-concerns

- ***Public concern:*** The public are concerned about the misuse of deep fakes, they are hard to detect and technology is advancing rapidly. The public have limited understanding, and there is a risk of public misinformation especially as the deep fakes become more sophisticated. It is good to look for inconsistencies when trying to decide if an image is a fake, such as mismatched earrings, inconsistent eye blinking etc.
- ***Worries and considerations:*** Deep fakes are increasingly being used for malicious purposes, such as the creation of pornography, and modern tools for creating them are readily available and increasing in sophistication yielding better and better results. Even though public awareness is increasing, the ability to detect a deep fake is not. However some recent research has shone a lens on who might be better at detecting them.
- ***Research by Ganna Pogrebna:*** Ganna is a decision theorist and a behavioural scientist working at the Turing Institute. She recently gave a talk by Zoom on her empirical study into "Temporal Evolution of Human Perceptions and Detection of Deep fakes". Ganna identified a range of personality traits (37) which could be measured using psychological measurement scales e.g. Anxiety, extraversion, self-esteem etc. Based on the description of the trait she then developed an algorithm. The hypothesis was based on the "big five" personality traits (openness, conscientiousness, extraversion, neuroticism and agreeableness).

The study commenced with a small group of 200 people, and has now increased to 3,000 people in each of five different Anglophone countries: UK, US, Canada, Australia, New Zealand. As Ganna has a large group of deep fakes (dataset) she can test using lots of different people not just using images of actors and politicians as in some studies. This has yielded a copious amounts of data, including cross sectional data from representative samples.

Continued…

Continued from previous page:

Each participant was subjected to 6 deep fake algorithm variations in a between subjects design.

- ***Results*** : People's ability to detect deep fakes gradually declines as the quality of the deep fakes improves. However, those people who show high emotional intelligence, conscientiousness and are prevention focused are better at detecting deep fakes. Neuroticism, resilience, empathy, impulsivity and risk aversion were traits coming in a close second with these people having better results. 2% of participants (which is low) were very good at detecting deep fakes (although no exact definition of "very good" was presented). They have three traits which are statistically scored higher than other participants: conscientiousness, emotional intelligence and prevention focus – they all detect well. General intelligence and knowing about technology does not make you able to detect deep fakes better, testing for general versus emotional intelligence could be an interesting addition to the data. It will be good to see the full results in terms of exact performance and effect size when published.
- ***So What:*** We are getting familiar with deep fakes and with talking about "hallucinations" such as content created by ChatGPT, these are assertions confidently made by algorithms even though they are far removed from the truth. The future technology is exciting, possibilities are endless with new technologies emerging at an exponential rate, but we need to question more than ever what we see and what we read.

Let us know what work you are doing in the deep fake arena – we'd love to hear from you – details on page 1

CAISS Bytes

Sir Paul McCartney has used AI to complete a Beatles song that was never finished. Using machine learning Sir Paul said they managed to "lift" the late John Lennon's voice and get the piece of work completed. By extracting elements of his voice from a "ropey little bit of a cassette", the 1978 song entitled "Now and Then" will hopefully be released later this year. "We had John's voice and a piano and he could separate them with AI. They tell the machine, 'That's the voice. This is a guitar. Lose the guitar'. This was not a "hard days night" and was certainly faster than "eight days a week", it will be interesting to hear the finished result and we may be "glad all over" that they did not "let it be".

Link here: https://www.bbc.co.uk/news/entertainment-arts-65881813

Using AI Chatbots to learn a language

The BBC have reported that students are switching to AI to learn a language. AI has benefits as it will not judge you if you make a mistake. With Spanish for example it can give regional variations such as Argentinian and Mexican Spanish. However, as useful as it can be for practising it can be repetitive, corrections are missing and words can be invented. As a supplement to other methods AI could have a place in cementing knowledge and making practice fun. Link here: **https://www.bbc.co.uk/news/business-65849104**

Continued from previous page:

Review of paper: "Fooled twice: People cannot detect deep fakes but think they can" – Nils C Kobis, Barbora Dolezalova & Ivan Soraperra

In this study the authors show that people cannot reliably detect deep fakes, even if they had their awareness raised and received a financial incentive, their detection accuracy was still poor. People appear to be biased towards mistaking deep fakes as authentic videos rather than the other way around and they also overestimate their detection abilities. Is seeing really believing?

These manipulated images, whilst entertaining can have a dark side. Large scale use of facial images are being used to create fake porn movies of both men and women which could impact their reputation; or in the case of a fake voice remove the life savings from someone. Calwell et al, 2020 ranked the malicious use of deep fakes as the number one emerging AI threat to consider.

This is an issue as the ability to create a deep fake using Generative Adversarial Networks (GANs) is not just in the realm of the experts but accessible to anyone, expert knowledge is not required. Extensive research in judgement and how people make decisions shows that people often use mental shortcuts (heuristics) when establishing the veracity of items online. This could, the authors posit, lead to people becoming oversensitive to online content and then fail to believe anything – even genuine authentic announcements by politicians. However, the counter argument is that fake videos are the exception to the rule and "seeing is believing" is still the dominant heuristic. This study tested both these competing biases – "liars divided versus seeing is believing".

The results of the study showed that people struggled to identify deep fake videos due to their inability to do so, not just that they were lacking in motivation. They also found that people were overly optimistic with a systematic bias exhibited towards guessing that the videos were authentic.

It could be argued that humans process moving visual information more effectively than other sensory data, results showed a slightly better than chance result and this is worse than when static images are used. Could this be due to inattention? More research is needed in this area.

The authors also found two related biases in human deep fake detection, participants were told 50% of the videos were fake, but still 67.4% were deemed to be authentic, so this was not related to their ability to guess so not deliberate – they were using their judgement. The other bias was related to the "Dunning Kruger"* effect, people over estimated their ability to detect deep fakes, particularly low performers were over confident. Overall people did really think that "seeing is believing".

Conclusion – Deep fakes will undermine knowledge acquisition as our ability to detect them is not due to a lack of motivation but an inability to do so. The videos used in this study did not have an emotional content which may have yielded different results. More work is definitely needed in this area.

Link to the paper here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8602050/

Reference: Caldwell M., Andrews J.T., Tanay T., Griffin L.D. AI-enabled future crime. *Crime Sci.* 2020;9:1–13. [Google Scholar]

**The Dunning-Kruger effect occurs when a person's lack of knowledge and skills in a certain area cause them to overestimate their own competence.*

Continued from previous page:

Protecting World Leaders Against Deep Fakes by Shruti Agarwal & Hany Farid.
The authors have developed a forensic technique that models facial expressions and movements that typify an individuals speaking pattern. These correlations can be violated during the creation of a deep fake video and so can be used for authentication.
https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf

CAISS Out and About - Conferences coming up…..

CREST international conference on Behavioural and Social Sciences in Security (BASS23), University of Bath, UK, from 11-13th July 2023.
The conference themes are Risk Assessment and Management; Gathering Human Intelligence; and Deterrence and Disruption.
Link here: https://crestresearch.ac.uk/bass23/

Look out for CAISS: Can AI Really Predict Political Affiliation? – Matthew Asher giving a Lightening talk on 12th July in the 14:00 and 15:00 session.

Abstract: Recent research suggests that it is possible to infer traits and behaviours—e.g., political leaning, criminality— from photos of people's faces (e.g., Wu & Zhang, 2016). Using profile images and AI technology researchers claim to have trained algorithms to classify people as conservative or liberal with 72% accuracy (Kosinski, 2021). Critics of this approach draw parallels with the unscientific practice of physiognomy and highlight risks of promoting discrimination. Despite the validity of these concerns little research exists examining this application of AI/machine learning—is it replicable and reliable? We sought to examine this question by gathering 1,998 facial images available online of liberal/conservative politicians. Following Kosinski's method, a convolutional neural network descriptor (VGG-Face2, Cao et al,. 2018) was used to extract a perceptually meaningful representation of each face which was then used to train a classification model. Our model achieved 66% accuracy: however, with further examination we found that 1) the model shows a bias to classify individuals as conservative and 2) around 25% of faces are never correctly classified. We are examining whether other factors (e.g., background colour) might better account for the model's accuracy than facial features and what this means for adopting AI for predicting behaviour.

A report to be included in the following issue on the
9th international Conference on Computational Social Science
which will be held in Copenhagen, Denmark on 17-20 July 2023
Link here: **https://www.ic2s2.org/**
This conference will bring together researchers from around the world in economics, sociology, political science, psychology, cognitive science, management, computer science, statistics and the full range of natural and applied sciences committed to understanding the social world through large-scale data and computation.