

Cody Gregg

Testing human ability to detect 'deepfake' images of human faces

The article “Testing human ability to detect ‘deepfake’ images of human faces” explains the threat of deepfakes and how people struggle to be able to tell what is a deep fake and what isn’t, the article goes into details on how they conducted test with to see if they could tell which images where deepfakes and which images wheren’t. Deepfakes are an evolving threat that has started to become more common in use for misinformation, targeting individuals, and even used for inappropriate entertainment.

The social science concerns of deepfakes has become more apparent through the years because of how easy it is to create it and the fact that people struggle with being able to tell if something is a deepfake. Phishing attacks are also starting to involve the use of deepfakes, making it harder for people to be able to tell if an email they got is fake. It also affects politics and how people see information about certain topics or elections; in the United States we are starting to see a lot of deepfake content being pushed out to the citizens by politicians and regular people. The way that people interact with deepfakes varies between different levels of knowledge on the subject, age, race, and gender; the older audience tends to fall for deepfakes more than the younger audience due to growing up more around the internet and being able to explore it.

In the article they outline the questions that they were seeking answers to during the tests. “Are participants able to differentiate between deepfake and images of real people”, a very straight forward question. “Do simple interventions improve participants’ deepfake detection accuracy”, the example that is given in the article is that they told the participants which images

were deepfakes and allowed them to study them before doing part of the experiment. “ Does participants' level of confidence align with their accuracy ”. The experiment was done by using 50 real images from Flickr that were randomly selected and 50 deepfake images that were created through StyleGAN2 which was trained using 70000 images from Flickr. They used 273 participants that they split into four groups; Control, Familiarization, One-time Advice, and Advice with reminders. Key data that was gathered was the mean age of participants being 26%, 39% of the participants were female. The full experiment breakdown for groups including real and fake images; Control 68.40% for real images (R) and 51.75% deepfake images (F), Familiarization 59.63% (R) with 62.01% (F), One-time advice 65.77% (R) with 62.25% (F), Advice with reminder 60.49% (R) with 69.10% (F), Overall 63.87% (R) with 60.99% (F). The experiment showed that people can tell that images are fake or real, but they still aren't nearly perfect at being able to tell which images are deepfakes. It also showed that people have doubts about real images, which can psychologically become a problem that can be taken advantage of later on.

In class we talk about how people interact with the internet and how the internet has an effect on people as it keeps evolving and adding more threats that can be targeted at people, deepfakes is evolving into a very serious thing that is making images become more real and learning off of images of people. Females tend to be the victim of deepfake attacks whether someone is making fake inappropriate images of them and even blackmailing them, the older generation tends to fall for misinformation created by deepfakes and tend to spread it to others; normal people in general can be victims of phishing attacks, catfishing, and even misinformation created all by deepfakes. This experiment shines light on the seriousness of deepfakes evolution

on the population and how people are starting to understand what is fake, it also shows that people tend to doubt stuff that is actually real which will start to have an affect on society.

Source

Testing human ability to detect 'deepfake' images of Human Faces | Journal of Cybersecurity | Oxford academic. (n.d.-b).

<https://academic.oup.com/cybersecurity/article/9/1/tyad011/7205694>