A Review of: Testing Human Ability to Detect 'Deepfake' Images of Human Faces

The topic of deepfake images and videos relate to the principles of social sciences by asking, what can you trust? As the interactions between humans and technology increase daily, so has the level of trust that humans have developed for technology. This trust is based upon the successful use of technology without deceit. Deepfakes undermine that trust by breaching the connection between what humans perceive and what technology can provide.

In the reviewed research paper by Sergi D. Bray, Shane D. Johnson, and Bennett Kleinberg, they investigate the human capacity to identify deepfake images generated by the StyleGAN2 algorithm. They also assess the effectiveness that simple interventions have on improving detection. Bray et al. (2022) describes deepfakes as, "Computationally created entities that falsely represent reality. They can take image, video, and audio modalities, and pose a threat to many areas of systems and societies, comprising a topic of interest to various aspects of cybersecurity and safety" (p. 1).

The research questions were: Are participants able to differentiate between deepfake and images of real people above chance levels? Do simple interventions improve participants' deepfake detection accuracy? Does a participant's self-reported level of confidence in their answer align with their accuracy at detecting deepfakes?

The research method utilized for this research paper was an online survey. The survey consisted of two-hundred-eighty participants. Each was shown twenty images that were randomly selected from fifty real images and fifty deep fake images. Participants were

instructed to identify the fake images and the real images and give reasoning behind the decision they made.

Data analysis for this research paper per Bray et al. (2022) utilized an a-priori statistical power analysis to estimate the sample size to detect a moderate effect size across the experimental conditions. A one-way analysis of variance was conducted for the power analysis. The Prolific online platform was used to recruit participants (p. 7).

The research paper relates to several topics discussed in class. Firstly, the psychological standpoint, making people believe that something is real simply because it looks real, can weaken the trust that people place in technology. Secondly, using deepfakes in social engineering may make people more susceptible to cyberattacks because of a false sense of trust. Lastly, using deepfake propaganda could be used to manipulate the public's opinion on certain subjects.

The topic of deepfake media could hinder marginalized groups by using deepfake media to alter the public's perception of that group. For example, social media could report fake crimes in an area using a respected reporter's appearance and voice. The implications of using deepfake and social media with no recourse could be damaging causing false accusations.

The overall contribution of the study to society was the realization that deepfakes are difficult to identify. The human eye, based on this research, was only able to detect sixty-two percent of deepfake media presented to participants. This leaves a huge gap

when deciphering what is real and fake on the internet. The studies' findings showcase a great need for solutions to detect deepfake media.

References

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2022). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity, 9*(1), 1-18. https://doi.org/10.1093/cybsec/tyad011