Testing Human Detection of Deepfake Images: A Cybersecurity and Social Science Review

Octavia Wade

Old Dominion University

CYSE 201S Cybersecurity and the Social Sciences

Professor Diwakar Yalpi

October 5, 2025

Introduction

I decided to pick the article "Testing human Ability to Detect 'Deepfake' Images of Human Faces," by researchers Bray, Johnson, and Kleinberg. This research talks about how well an individual can recognize that an image is AI-generated and whether we can make sources that can improve the detection of AI-generated photos. This topic of research emphasizes psychology, sociology, and cybersecurity standpoints and shows how a human vulnerability can be impacted by technologically mediated deception.

Relation to the social Sciences

In this research, it showcases the social science principles by focusing on perception and decision-making. Deepfakes represent a technical innovation but also it raises questions about trust, misconception, and the manipulation between social reality and AI reality. By analyzing how people judge authenticity, researchers can determine how exactly technology is influencing human cognition and social behavior towards technology (Bray et al., 2023).

Research Focus and Variables

The researchers ask three questions that center on the topic they are researching. They asked: whether participants could identify deepfakes if it was present to them, whether there are interventions out there that could improve detection, and whether a participants' confidence will help with accuracy. They hypothesized accurate data, benefits from interventions, and how a positive attitude can help with confidence and correct information. The independent variable in the article was based on the type of intervention, while the dependent variables only were there to do detection on accurate and self-reported confidence results (Bray et al., 2023).

Methods, Data, And Analysis

The researchers decided to employ an online experiment with over 280 participants who were randomly assigned to one of the four groups that they formed. Each of the participants were judged based on 20 random images that were drawn from a pool of 50 authentic original photos and 50 StyleGAN2-generated deepfakes photos. In their data, it included categorical response, rates of the participants' confidence, and short written explanations. Their analysis used descriptive statistics, a t-test, and ANOVA to compare results from different conditions. They found an average accuracy of 62%, which is slightly above the chance. However, there was no significant improvement for any invention. Meanwhile, the confidence level of the participants

still remained overly high regardless of what the correctness stated on the results, this suggested that it was a troubling in the disconnect between perception and reality (Bray et al., 2023)

Connection to other Course Concepts

The article I choose can relate to modules 2 and 3. In Module 2, we talk about principles such as empiricism, skepticism, and determinism. Detecting deepfake can demonstrate empiricism because it reflects on the observable and measurable data, such as the participants' using their judgment and confidence. It also embodies the principle of skepticism due to the challenges of assumptions that people can easily spot out the manipulation of social media. Lastly, the principle of determinism can be applied cause of the exposure to misinformation or the biases mindset that impacted an individual decision-making process.

In Module 3, we talked about different types of experiments that are used in cybersecurity research. Bray et al. (2023) did some of the experiments we talked about in class; for example, they conduct experiments by randomly assigning participants into intervention conditions and recording their response from across different tasks. This aligns with the article and Module 3 because it talks about how researchers use these experiments to test whether or not their intervention can change people behavior, specifically if humans can get familiar and improve the detection of deepfake photography. In their study, they reflected on one challenge that was talked about in Module 3, experiments in cybersecurity are rarely seen as mirror classic lab experiments due to having difficulties such as the mix of random samples and the real-world effects. Bray et al, (2023) online survey-based experiment shows how researchers are able to adapt to methods that surround cyberspace but still maintaining a scientific rigor.

Marginalized Groups and Societal Concerns

Deepfakes had a huge threaten impact on marginalized groups, specifically on women and minorities groups. They were frequently targeted by getting harassed, image-based abuse, and campaigns that were very unsettling. Documents showed how difficult it was to detect theses deepfakes and how studies went underscores how deepfake technology made it worse to exist inequalities. With this, it highlights why the need for protective measures that can go beyond the technical solutions and address the borderline issue of justice and representation (Bray et al., 2023).

Contributions to Society

How this article contributes to society, well it reveals how limited the human detection of deepfakes and the ineffectiveness of simple training interventions can do. With a caution against an overreliance on what the human judgment when it comes to politics, online dating, and financial security. In their study, they discovered that inform policymakers, educators, and cybersecurity analyst are all under the stress due to the response to the deepfakes that are combining the technological tools with the education and awareness campaigns (Bray et al., 2023).

Conclusion

Overall, Bray and other researchers (2023) provided an important straightforward examination on how human vulnerability can be impacted by deepfakes. They found cognitive limits on an individual's thinking, but also it highlighted the broader societal risks it can have on synthetic media later in the future. This draws attention to both the technical and social aspects of it all. Their study undergoes the urgency of having an interdisciplinary approach to minimizing the harms of digital deception that is polluting society’s mind and social media.

References

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, *9*(1). https://doi.org/10.1093/cybsec/tyad011