Article Review #2: Testing Human Ability to Detect 'Deepfake' Images of Human face

Student Name: DESTIN DANQAH

School of Cybersecurity, Old Dominion University

CYSE 201S: Cybersecurity and the Social Sciences

Instructor Name: PROFESSOR DIWAKAR YALPI

Date: November 14th, 2025

Introduction/BLUF

During the 2023 article, Bray, Johnson, and Kleinberg investigate how well human can see the difference between AI- generated "deepfake" pictures of human faces from real ones and whether simple creations improve accuracy. The bottom line: participants performed only marginally above chance out of all the deepfake vs. real images people were asked to judge, they only got about 62% of them correct, confidence in judgements was high and unrelated to accuracy, and none of the simple interventions significantly improve performance.

Relation/Connection to Social Science Principles

This study links strongly to social science principles concerned with human decision making, trust, perceptions of authenticity, and social engineering.

Key principles include:

Human Behavior and Cognition: how individuals process visual stimuli and make judgements about authenticity.

Trust and risk perception: participants think their judgements were accurate regardless of weak actual performance, reflecting overconfidence bias.

Social Interaction and Deception: deepfakes represents new ways of deceiving people in the digital environment.

Technology and Society: the study shows how AI driven media can undermine human abelites to detect deception, with implications for cyber society.

Equity and access: while not the man focus, the findings suggest that all individuals are vulnerable to this type of manipulation, no matter what skill levels you may have.

Systems and institutions: the study tells us that the need for institutional responses rather than relying mainly on individual detection.

Behavioral changes: the failure of interventions shows that behavioral change is hard, and simple advisory steps may not be enough to mitigate the threat.

Research Question /Hypothesis/ Independent Variable/Dependent Variable

- Research Question:

1. Do simple interventions improve deepfake detection accuracy?
2. Is participants self reported confidence aligned with their accuracy?
3. Are participants able to differentiate between deepfake and real human face images above chance level?

- **Hypothesis**: The author hypothesize that participants would perform above chance, that intervention group would outperform control, and that higher confidence would correspond to higher accuracy.
- **Independent Variable**: Experimental condition
- **Dependent Variable:** (a) participant accuracy when labeling images (percentage correct), (b) self – reported confidence level for each judgement and (c) coded use of "tell tale" visual features in participants 'reasoning'.

Types of Research Methods used

This is a quantitative experimental study using online survey methodology. The study had approximately 280 participants were randomly assigned t one of four conditions and asked to evaluate images and provide confidence and reasoning.

Types of Data Analysis used

Data had per participant accuracy, confidence ratings, and free texting reasoning coded for use of telltale features. Analysis techniques included one sample t-tests, one way ANOVA for condition differences, repeated measures ANOVA for image type x condition interactions, and coding of textual reasoning with NVivo

Connections to other Course Concepts

This article restarts concepts we have covered like social engineering like deepfakes are a new vector of manipulation, trust exploitation like high confidence despite low accuracy, and cyberthreat modelling for example deepfakes represent an emerging threat. It also highlights the limitation of human centric defenses and the need for layered security strategies, including technology, education, and institutional controls.

Connections to the Concerns or contributions of Marginalized Groups

Although not the main focus, the study's finding has implications for marginalized groups who may have less training, less resources, or lower digital literacy. These groups may be more vulnerable to deepfake based deception because human detection has already started to be weak. The authors note that deepfakes may amplify existing harms like harassment of women and minority online.

Overall societal contributions of the study/Conclusion

In conclusion, this study makes a significant contribution by empirically showing how poorly humans perform at detecting deepfake face images, even when shown simple guidance, and by highlighting a crucial gap in individual level defenses. The findings suggest that society cannot rely on individuals alone to defend against deepfake threats; institution, education, regulation, and technology must intervene. For cybersecurity policy through a social science framework, this research underscores the importance of behavioral research and systemic responses to emerging AI theats.

Reference

Bray, Sergi D, et al. “Testing Human Ability to Detect “Deepfake” Images of Human Faces.” *Journal of Cybersecurity*, vol. 9, no. 1, 1 Jan. 2023, academic.oup.com/cybersecurity/article/9/1/tyad011/7205694?searchresult=1, https://doi.org/10.1093/cybsec/tyad011.

Article Link:

[https://academic.oup.com/cybersecurity/article/9/1/tyad011/7205694?searchresult=1]