Jimmy Benjamin

Professor Diwakar Yalpi

CYSE 201S

11/16/2025

Article Review #2 – Testing Human Ability to Detect ‘Deepfake’ Images of Human Faces

Introduction

This article examines whether individuals can reliably detect AI-generated deepfake images and whether simple interventions improve detection accuracy. The topic holds strong relevance for social science principles such as human perception, cognitive bias, and digital trust. By evaluating how people interpret and judge online information, the study reinforces how human behavior and decision-making intersect with cybersecurity risks.

Research Questions, Hypotheses, and Variables

The study explored three core questions: (1) Can individuals identify deepfakes above chance? (2) Do brief interventions improve accuracy? (3) Does confidence align with actual performance? The authors hypothesized that participants would perform better than chance, that interventions would improve accuracy, and that confidence would correlate with

accuracy.

The independent variable was the type of intervention: control, familiarization, or advice. The dependent variables included detection accuracy, confidence ratings, and qualitative reasoning statements.

Methods, Data, and Analysis

Researchers conducted a controlled online experiment with 280 participants randomly assigned to one of four groups. Each participant reviewed 20 images and labeled them as real or AI-generated, provided a confidence score, and explained their reasoning.

Data included quantitative correctness scores, confidence ratings, and qualitative explanations. Analysis revealed an average accuracy of roughly 62 percent. Interventions did not significantly improve performance. Confidence showed weak correlation with actual accuracy.

Relation to Course Concepts

The article reinforces core social science concepts covered in class, including cognitive bias, overconfidence, empirical research design, and operationalization of variables. It also aligns with discussions on technological trust, digital literacy, and human-technology interaction—demonstrating how people form judgments under uncertainty and how these

processes impact cybersecurity outcomes.

Impact on Marginalized Groups

The findings highlight concerns for communities with lower digital literacy, who may be disproportionately vulnerable to deepfake-driven misinformation or identity manipulation. Overconfidence despite low accuracy increases susceptibility to deception. The article emphasizes the need for accessible media-literacy education and systemic protections to prevent technology-driven inequities from widening.

Contributions to Society

The study provides evidence of limitations in human detection of deepfakes and challenges assumptions about individuals' ability to self-protect online. By identifying gaps in judgment and intervention effectiveness, it underscores the importance of technical safeguards, public-policy responses, and educational strategies. The work contributes to societal understanding of synthetic media risks and the need for multi-layered cybersecurity solutions.

Conclusion

This article effectively integrates social science principles with cybersecurity challenges by examining how people perceive deepfake images and how interventions influence judgment. The research demonstrates key gaps in human accuracy and confidence, reinforces the

importance of media literacy, and highlights risks that disproportionately affect marginalized groups. Its findings contribute to broader societal efforts to address synthetic media threats through coordinated technical, educational, and policy-based approaches.

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), tyad011. https://doi.org/10.1093/cybsec/tyad011