Article Review #2: Testing human ability to detect ‘deepfake’ images of human images

Student Name: Kennice Allea Balmoria

School of Cybersecurity, Old Dominion University

CYSE 201S: Cybersecurity and Social Sciences

Instructor Name: Diwakar Yalpi

Date: 06 April 2026

Introduction

The article *Testing human ability to detect 'deepfake' images of human faces* examines how well people can distinguish AI-generated deepfake images from real human faces, and whether simple interventions can improve accuracy. This topic is deeply connected to the social sciences because it explores human perception, decision-making, trust, and vulnerability within digital environments. As deepfakes become more common, understanding how people interpret visual information is essential for public safety, cybersecurity, and social well-being.

Relation to Social Science Principles

This study is directly related to core social science principles such as:

- **Relativism:** In this study, participants relied on subjective cues such as "naturalness," symmetry, or emotional expression, when deciding whether an image was real. The authors note that participants highlighted different parts of the same image and gave different explanations for their decisions, showing that judgments varied widely across individuals.
- **Determinism:** The study demonstrates this by showing that accuracy was influenced by factors such as image quality, perceptual imperfections in StyleGAN2 outputs, and participants', suggesting that specific visual features causally shaped detection ability.
- **Objectivity:** The researchers used randomized assignment to four groups, standardized image sets (50 real, 50 deepfake), and statistical tests like t-tests and ANOVA to ensure neutral measurement. They also coded qualitative responses systematically to avoid subjective interpretation. Participants' confidence was

analyzed objectively as well, revealing that confidence remained high regardless of correctness.

- **Parsimony:** Despite testing three different interventions, the authors conclude that humans are only slightly above chance (62%) at detecting deepfakes and that simple advice does not significantly improve accuracy. This straightforward explanation avoids unnecessary complexity and supports the idea that human detection alone is insufficient.
- **Empiricism:** The study embodies this principle by collecting measurable data: accuracy scores, confidence ratings, and per-image performance. For example, accuracy ranged from 85% on some images to below 30% on others, and one in every five images had below-chance accuracy. These findings were grounded in observable patterns rather than assumptions.
- **Skepticism:** The study challenges the common belief that people can "just tell" when something is fake. Participants reported high confidence was "high and unrelated to accuracy," revealing a dangerous mismatch between perception and reality.
- **Ethical Neutrality:** Although deepfakes have harmful uses, including fraud, harassment, and political manipulation, the authors analyze detection ability without blaming participants for being deceived. They focus on understanding the human condition and vulnerability rather than assigning fault.

Research Questions, Hypotheses, Independent & Dependent Variables

- **Research Questions:**
    1. Are participants able to differentiate deepfake images from real images above chance level?
    2. Do simple interventions (familiarization or advice) improve detection accuracy?
    3. Does participants' confidence align with their actual accuracy?
- **Hypotheses (Implied):**
    1. Participants will perform slightly above chance (50%) when identifying deepfakes.
    2. Interventions will improve accuracy.
    3. Confidence will correlate with accuracy.
- **Independent Variables:**
    1. Experimental condition (Control Familiarization, One-Time Advice, Advice with reminders).
    2. Image type (real vs. deepfake)
- **Dependent Variables**
    1. Accuracy of labeling images as real or AI-generated.
    2. Self-reported confidence levels.
    3. Qualitative reasoning participants provided.

Research Methods Used

The study used a **quantitative experimental design** with four randomly assigned groups. Participants completed an outline survey where they viewed 20 images and judged whether each

was real or AI-generated. The interventions varied by group, allowing the researchers to test causal effects.

The study also incorporated **qualitative elements**, such as open-ended responses explaining participants' reasoning, and a grid-based tool where participants highlighted parts of the image that influenced their decisions.

This mixed-methods approach strengthened the analysis by combining numerical accuracy with insights into the human thought processes.

Types of Data and Analysis

The authors collected:

Quantitative data:

- Accuracy scores
- Confidence ratings
- Statistical comparisons across conditions
- Per-image accuracy patterns

Qualitative data:

- Written explanations
- Highlighted image regions

Analytical techniques included:

- One-sample t-tests to compare accuracy to chance
- ANOVA to compare accuracy across conditions

- Correlation analysis between confidence and accuracy
- Coding of qualitative responses to identify patterns in reasoning

The results showed that participants averaged 62% accuracy, only slightly above chance, and none of the interventions significantly improved performance. Confidence remained high, regardless of correctness, revealing a concerning mismatch between perception and reality.

Connections to Course Concepts

- **Misinformation and disinformation:** Deepfakes amplify the spread of false content and undermine trust.
- **Cognitive biases:** Overconfidence bias was evident; participants felt sure even when wrong.
- **Media literacy:** The study highlights the need for stronger digital literacy skills.
- **Technology and inequality:** Access to knowledge about digital threats varies across populations.
- **Social construction of reality:** Deepfakes challenge our ability to rely on visual evidence.

These themes reinforce the importance of understanding how technology shapes human behavior and social systems.

Relation to Marginalized Groups

Deepfakes disproportionately affect marginalized groups in several ways:

- **Women and minorities face higher rates of harassment,** including non-consensual deepfake pornography and targeted abuse.
- **Political manipulation often targets vulnerable communities,** spreading false narratives that influence public opinion.
- **Fraud schemes,** such as romance scams, frequently target older adults or socially isolated individuals.
- **Bias in AI training data** can make certain groups more vulnerable to misrepresentation or exploitation.

The articles acknowledge these broader societal risks, emphasizing inequalities and exposing marginalized groups to new forms of harm.

Overall Contribution to Society

The study makes a clear contribution by showing that people are only slightly better than chance at detecting deep-fake images, even when given simple guidance. This finding highlights the need for stronger digital literacy, better technological safeguards, and more effective public education about AI-generated media. By revealing how easily individuals can be misled, and how confident they remain despite being wrong, the research provides important evidence for policymakers, educators, and cybersecurity professionals working to protect the public from digital deception.

Reference

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). *Testing human ability to detect 'deepfake' images of human faces.* Journal of Cybersecurity, 9(1), 1-18. https://doi.org/10.1093/cybsec/tyad011

Article Link:

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