

Article Review One: Testing human ability to detect 'deepfake' images of human faces

Thomas Gable

Department of Cybersecurity, Old Dominion University

CYSE 201s: Cybersecurity and the Social Sciences

Diwakar Yalpi

September 19, 2025

Introduction

“Testing human ability to detect ‘deepfake’ images of human faces” (Bray et al., 2023) examined humans' ability to discern deepfake images from real ones, looked at whether giving any prior detection knowledge could help, and if their confidence level was consistent with their detection accuracy. The authors started this article by giving some background knowledge, highlighting the growing concern of AI-generated images as they have become increasingly accessible to the public for use. They go into detail about the different details surrounding deepfakes, such as they can come in multiple media forms, their use in real-world applications both positively and negatively, and how the public is becoming more and more distrusting of the media. The text gives evidence for their concern, stating “For example, in a recent study on ‘AI and Future Crime’ [46], 31 representatives with expertise in AI from academia, policing, government, the private sector, and state security agencies, rated deepfake technologies to be the top-ranking threat associated with AI” (Bray et al., 2023). Furthermore, they dove into technical solutions, such as detection algorithms, but ultimately found they weren’t practical and had too many flaws to be effective. They then argue that these artificial images contain obvious fake visual aspects that could aid individuals in detecting them if properly educated. This leads them to their study, which examines exactly this.

Research Questions, IV, and DV

The authors created three main research questions: “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?” (Bray et al., 2023). They hypothesized that participants would be able to accurately discern deepfakes from real images above a 50% rate, that providing the groups with knowledge would increase detection accuracy, and a person's confidence would correlate with their accuracy level. The independent

variable was the varying levels of knowledge presented about deepfakes. The dependent variable was their accuracy in classifying images as deepfake or real.

Research Methods

The authors found people on a social media platform, offering them hourly money, with extra for good performance, to participate. They were then assigned to four groups, with the first one being a control group who were just shown the images. Group number two was given twenty deepfake image examples and was able to examine them prior to the questionnaire. Group number three was given some explicit revealing attributes to look for, which are commonly present in deepfakes, paired with two images that contained them. Group four was the exact same as group three, but was displayed this information under each question on the survey, whereas group three only saw it before. Then, the experiment was deployed on a web server, where each participant could type in the questionnaire URL, read about the study, and answer the questions. The individuals were shown 20 images, one at a time, with a question asking for their classification (real or deepfake), a ten-point scale to rate their confidence, a textbox to explain their reasoning, and a box grid over the picture, where they could click on the area they detected the sign. This study used both qualitative and quantitative data, as shown above.

Results and Data Analysis

The participants answered correctly 60 percent of the time overall, with no notable difference in the groups, indicating that the educational interventions did not increase participants' accuracy levels. It is important to note that groups two, three, and four got a little better at spotting deepfakes, but worse with real ones, cancelling each other out. The authors put the free-form answers through a system that looked for keywords that mentioned the provided indicators. It was found that individuals who were exposed to the information, groups three and four, used it when making their decisions, more so than the other two. But this caused individuals in groups three and four to wrongly classify real images as deepfakes using the

cues, more often, explaining why the accuracy was consistent across the board, as they were just classifying more as AI-generated overall. Furthermore, it was found that some of the signs correlated with better accuracy levels than others, and were therefore, more reliable. The authors also found that confidence levels were consistently on the higher end overall, no matter what group they were in. Furthermore, it was discovered that those who felt more confident in their answer were not more likely to answer correctly.

Conclusion and Societal Contributions

Overall, it was found that informing individuals of specific attributes that are found in a deepfake, but not a real image, doesn't have a significant effect on their ability to distinguish between the two. This is a concern and should be further investigated to find a solution because your average person is only able to detect AI-generated images at about a 60% rate. These images can have a harmful effect on the public, creating realistic pictures that can be used to represent and formulate malicious narratives. This study also showed that participants demonstrated a high confidence in their answers even when they were wrong. This finding is concerning because this shows that in the real world, people may be seeing deepfakes and are sure that they are real, leading them to believe the content, when it may be false. The findings in this study could be used to further investigate the cues that were found to be more effective to see if understanding those can lead to increased detection accuracy.

Social Science Principle Relation

The authors of this article followed the Ethical Neutrality principle, as they explicitly stated they kept participants' data anonymous and got permission from the Departmental Ethics Committee to carry out this experiment. They also followed the principle of Determinism, as their study was based on examining whether individuals being exposed to more knowledge before answering questions would cause them to be better at detecting deepfakes. Lastly, the study followed the principle of Objectivity, as they attempted to be as fair as possible in their data analysis. Even though the results didn't align with their hypothesis or desires, they still

objectively looked at the data and came to a conclusion, which again denied what they previously believed.

Article Relation to Class Concepts

On top of already discussing three principles of social sciences, the strategy used to conduct the study, and the different aspects (IV, DV, research question, etc..), I will lastly talk about its relation to the concept of human factors. This experiment sought to examine whether educating individuals on the common revealing signs that an image is a deepfake would increase their ability to detect them. Furthermore, through this, it was able to assess the effectiveness of these certain signs, with some being used more correctly than the others. The goal of this study was to assess if deploying this information to the masses would actually aid in detecting these deceiving, harmful images because technical measures aren't cutting it. Even though the results weren't all that convincing, this could still lead to more coordinated and better optimized educational teachings using the cues that were shown to be more effective. This could invite more research and possibly lead to a widespread information campaign to educate the masses on the tell-tale signs to detect these deepfake images.

Relation to the Challenges, Concerns, and Contributions of Marginalized Groups

Deepfakes are notorious for affecting women by editing or creating inappropriate and explicit images, and sometimes videos of them. They are commonly posted on social media or sometimes even sold on a website for public consumption. This can gravely tarnish a female's public image, humiliating them and losing them potential opportunities like sponsors.

AI-generated images can also be used to amplify racial stereotypes, painting a certain ethnicity in a prejudiced manner, such as in a criminal context. This can be used to curate fake perceptions of these groups, possibly encouraging harsh judgment and discrimination. This study aimed to create a fix to this widespread, nasty issue by assessing whether simply enlightening individuals on the telltale signs of deepfakes was effective. If people are able to

detect the artificial media behind the fake narratives they create, they can be less manipulated and deceived.

Reference

1. Sergi D Bray, Shane D Johnson, Bennett Kleinberg, Testing human ability to detect 'deepfake' images of human faces, Journal of Cybersecurity, Volume 9, Issue 1, 2023, tyad011, <https://doi.org/10.1093/cybsec/tyad011>