The Threat of Deepfakes: An Overview of Human Ability to Detect Them

Carl Lochstampfor Jr.

Department of Cybersecurity, Old Dominion University,

CYSE 201S: Cybersecurity & Social Science, Professor Trinity Woodbury

February 21, 2025

Article Review Source:

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 1-18.

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

How does the topic relate to the principles of the social sciences?

'Deepfakes' are considered 'synthetic media:' they computationally create entities that falsely represent reality, taking different media forms while posing a serious cybersecurity threat to networks, systems, and societies. World leaders, high-profile academic researchers, and institutions identify deepfake technologies as a top-ranking AI threat (Bray, Johnson, & Kleinberg, 2023, p. 3). Crime applications include fraud like the 'grandparent scam,' authentication forgery to access secure systems and classified information, and fake video evidence of public figures speaking or acting in specific ways to manipulate support. A critical element of the deepfake threat is that it can deceive "a human actor into believing that it correctly represents reality, rather than merely (and explicitly) mimicking it" (Bray, Johnson, & Kleinberg, 2023). The fake material can circulate through many uncontrolled routes, creating new crime vectors, and people's trained response to it may be the only effective defense against it.

At the time of the study and after extensive research, the researchers agreed there were no technical solutions to fully address the deepfake threat sufficiently (Bray, Johnson, & Kleinberg, 2023, pp. 3-4). Thus, the researchers believed that humans are relatively reliable at labeling/identifying actual video stimuli, even in the presence of deepfake stimuli. Studying human performance can assist with understanding how to identify deepfake stimuli vs. actual video stimuli better and use that knowledge to train the public. The researcher's goal was to help reduce the risk of people falling prey to the predators behind the fake media.

What were the study's research questions and hypotheses?

The study's hypotheses and research questions were as follows:

1. Can participants differentiate between deepfake and images of real people above chance levels?
2. Do simple interventions improve participants' deepfake detection accuracy?
3. Does a participant's self-reported confidence level in their answer align with their accuracy at detecting deepfakes? (Bray, Johnson, & Kleinberg, 2023, p. 7).

What types of research methods were used? What types of data and analysis were done?

The researchers used several social science research methods, including archival research, cyberspace field studies, and experimental studies. First, to ensure the integrity and randomness of the image dataset, the researchers selected 50 authentic and 50 deepfake images from the FFHQ dataset. The deepfake data source StyleGAN2: FFHQ—created by NVIDIA's research team—is publicly available, open-source, and highly regarded in the academic community. According to Bray, Johnson, and Kleinberg (2023, p. 8), the StyleGan2 deep learning algorithm generated the

deepfake images "trained using 70,000 images from the FFHQ dataset. All deepfake images were collected without any curation element, representing the random output of the Style- GAN2 algorithm."

Second, the participants were selected and recruited through an online platform. The participant's median age was mid-twenties; more than half were male and from different nationalities. Third, the experiment was implemented using a web application written in Django and hosted on a secure server in the Netherlands (sergibot, 2022). Fourth, the researchers provided the participants with instructions on how to complete the experimental task. The researchers established a baseline control condition and three 'experimental' conditions. The researchers provided intervention content to the experimental groups to determine whether the advice would help or worsen the participants' decision-making process when choosing between deepfake and non-deepfake images.

Fifth, using a set of 20 trials, the analysis had participants labeling a stimulus as AI-generated or not based on performance and not chance while reducing the risk of cognitive overload. The experiment application consistently displayed images at a large resolution, one at a time. Lastly, during the experiment, participants were asked to provide a 'free text' explanation "to their reasoning behind their decisions and to clock on locations within the images that informed their choices" (Bray, Johnson, & Kleinberg, 2023, p. 9).

How does the topic relate to the challenges, concerns, and contributions of marginalized groups?

One of the "chief limitations of the study is that participants were mostly young, with a mean age of 26" (Bray, Johnson, & Kleinberg, 2023, p. 14). The other participant identification factors like sex, nationality, and first language had "no impact on the comprehension of the

experimental task or any elements of the survey…. There were no differences in any of these characteristics across conditions" (Bray, Johnson, & Kleinberg, 2023). However, the researchers agreed older targets are more often targets of fraud (such as the 'grandparent scam'), likely to be less aware of technological developments, and can have worse eyesight and attention to detail than younger targets (Age UK, 2019; Age UK, 2023; FBI, 2012; FBI, 2024). With elder fraud on the rise, it appears the elderly will be less accurate than their younger counterpart participants and fall victim to precipitation without some intervention, training, and/or education (FBI, 2024).

Lastly, a big concern was that the study showed marginal improvements in the participants' performance when given the intervention advice with the resolution's realism. Images in real-life scenarios come in different sizes, shapes, and resolutions. The smaller the size, the more difficult it is for the viewer to identify possible nuisances associated with deepfakes. Also, concerning the overall performance, none of the intervention conditions significantly improved participants' accuracy at detecting deepfakes from non-deepfake images (Bray, Johnson, & Kleinberg, 2023, 9). That implies that some professional assistance with handling deepfakes vs. real-media forms is insufficient and needs reexamination to have a helpful impact on society. Thus, deepfake detection accuracy and education for different age groups should be considered for future work to understand how groups like minors and older people can identify deepfakes versus non-deepfake media, including images, audio recordings, robocalls, and phishing emails.

How do concepts discussed in class relate to the article? What are the overall contributions of the studies to society?

First, the study seeks to understand the human factors and their ability to "detect deepfake images from similar authentic images, and into the methods by which they do so" (Bray, Johnson, & Kleinberg, 2023, p. 7). Identifying deepfakes and protecting oneself from their deceptive

practices affects cybersecurity and every sector of our societal system. That includes systems of education, politics and news media, criminal justice, infrastructure like physical buildings and transportation, and health care.

Second, researchers conducted a thorough power analysis before (a priori) collecting data. Bray, Johnson, & Kleinberg (2023, p. 7) wanted "to determine the appropriate sample size needed to detect an effect of a given size with a specified level of statistical power. The chosen sample size was justified based on statistical reasoning rather than being determined arbitrarily or after data collection (post hoc)". Thus, the researchers, prior to conducting the researching and obtaining the results, wanted to ensure the study had enough power beforehand to detect meaningful effects while reducing the risk of errors and biases.

Third, the study has a "strong focus on ecological validity and relevance to the wider context of the criminal misuse (e.g., in cases of fraud, such as dating scams) of the technology and policy implications" (Bray, Johnson, & Kleinberg 2023, p. 7). In other words, the researchers designed the empirical experiment with cognitive behavioralism in mind: to emulate realistic scenarios of people viewing different media images while behaving/participating in normal day-to-day activities like internet browsing or watching television at home, the office, library, or local sports bar.

Lastly, "the study delves deeper into participant decision-making, asking why participants were confident in making some decisions over others and objectively tracking whether or not that confidence had a relatively positive or negative impact on their results" (Bray, Johnson, & Kleinberg, 2023, p. 7). For example, in-text responses from the participants provided context about why they believed and were confident that some images were deepfake images while others were non-deepfake images.

References

Age UK. (2019, July 30). *Older person becomes victim of fraud every 40 seconds*. Retrieved from Age UK: https://www.ageuk.org.uk/latest-press/articles/2019/july/older-person-becomes-fraud-victim-every-40-seconds/

Age UK. (2023, June 7). *Older People, Fraud and Scams*. Retrieved from Age UK: https://web.archive.org/web/20230228071942/https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/reports-and-briefings/safe-at-home/rb_oct17_scams_party_conference_paper_nocrops.pdf

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 1-18.

FBI. (2012, April 2). *The Grandparent Scam: Don't Let It Happen to You*. Retrieved from Federal Bureau of Investigation (FBI): https://web.archive.org/web/20230321235226/https://www.fbi.gov/news/stories/the-grandparent-scam

FBI. (2024, April 30). *Elder Fraud, in Focus*. Retrieved from Federal Bureau of Investigation (FBI): https://www.fbi.gov/news/stories/elder-fraud-in-focus

sergibot. (2022, December 19). *FakeFacesSurvey*. Retrieved from GitHub: https://github.com/sergibot/FakeFacesSurvey