NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

IS6799 CRITICAL INQUIRY

Research Proposal DG-01-02

Supervisor: Dion Goh

Submitted by:

Name	**Matric No.**	**Email**
Guo Zhengxi		
Song Siyi		
Zeng Ruoyao		

Date: 2023/01/27

Aims of the research

This study aims to provide an understanding of perceptions of Deepfakes from the perspective of audiences, exploring if there are any strategies for Deepfake identification using a diary study in order to help audiences better discern Deepfake videos on the internet.

Background to the research

As technology developed rapidly, social media became a platform that significantly influenced daily human life. With the enormous amount of information emerging, plenty of fake news (FN) is created. Olan et al. (2022) mentioned in their article that fake news keeps altering people's point of view on critical issues and their societal values. Deepfakes could be one of the most popular technologies for people to create fake news.

The word 'Deepfakes,' as Nguyen et al. (2022) described in the article, is a combination of two words: 'deep learning' and 'fake,' in which people use AI technologies to swap the faces of celebrities or politicians to bodies in illegal images and videos. Deepfake technology will help generate a humorous or political video of a person saying anything without the consent of the person, whose appearance and voice are involved (Westerlund, 2019). For example, During President Trump's election campaign, his competitors used Deepfake to create some peachy news about him, which used to discredit him and make people distrust him. According to Köbis et al. (2021) talking about people's perception of Deepfakes, their research result shows that people can no longer detect Deepfakes since many of them are overconfident in their detection abilities. Recent studies related to Deepfake mostly rely on the technological surface, which focuses on how to create Deepfake videos but ignores the humanity part. In this research, our primary goal is to analyze the underlying reasons that influence people to perceive Deepfake videos. In addition, we will provide future strategies for helping people to authenticate Deepfake videos.

Objectives of the Research

1. Record participants' data to determine if there are any improvements in their Deepfake video detection ability in a period of time and record their individual differences.
2. Provide future strategies for helping people to authenticate Deepfake videos.

Literature Review

Deepfake background

Deepfake was created by a person who used the username "Deepfake", a combination of the words "Deep learning" and "Fake". This user publicly posted the video he made on the Internet, replacing female celebrities' faces in an adult porn video. (Albahar & Almalk, 2019). While Gamage and other researchers define Deepfake as “the application of deep learning methods to generate fake content-usually video, images, audio or text (Gamage et al., 2022, p. 2)”. This definition is a simple explanation of Deepfake technology in a broad sense. Other researchers, on the other hand, focus more on aspects related to the principles of the technology. In their research paper, Li et al. describe how the technique works as “the faces of a target individual are replaced by the faces of a donor individual synthesized by DNN models, retaining the target's facial expressions and head poses (Li et al., 2022, p. 72)”. This means that the technique relies on a large amount of data and algorithmic procedures to support it. All three articles provide a basic definition of the deep forgery technique from different perspectives, which will give our project a more comprehensive understanding of the technique and help us provide technical explanations to the participants in the subsequent phases.

Kietzmann et al. provide a more detailed description of the technical principles of deep forgery video generation. The technique uses an auto-encoder and a "Generative Adversarial Network" (GAN) machine learning model to merge and superimpose images or videos onto the source image or video, and a large sample of individuals' voices, facial expressions, and body movements are stitched together into fake content with the help of “Deep Neural Network (DNN) (Kietzmann et al., 2019). In simple terms, a database containing hundreds or even thousands of photos of the target task is placed in the algorithmic program for learning, which results in image replacement in the same encoder. This article provides examples in some detail to explain how Deepfake works and can help provide a technical knowledge base for our project analysis.

Impacts and threats of Deepfakes

Being a pioneering technological innovation, the possible negative effects of Deepfakes are considerable. Caldwell et al. (2020) categorized and rated the potential criminal and terrorist threats to AI technology. Among them, Deepfake is classified as high risk with the greatest concerns.

Dominated by the notion that seeing is believing, people trust video content more in terms of visual input (Frenda et al., 2013, Sundar, 2008). Therefore, Deepfake videos are more likely to be trusted as real (Köbis et al., 2021). Moreover, the threshold for participation in Deepfake is very low. With the help of Deepfake applications, such as FaceApp, Deepfake technology is widely accessible to the public (Köbis et al., 2021). Creating Deepfake videos is also becoming accessible to the general public without large training datasets (Nirkin et al., 2019).

Pornography is a widely used area of Deepfake, and the production of related fake videos brings many negative effects. A study conducted in 2019 points out that 96% of Deepfake videos are classified as pornographic (Ajder et al., 2019). A large number of pornographic Deepfake videos threaten people's right to reputation (Köbis et al., 2021). And the complex regulatory situation caused by Deepfake also demands a higher level of privacy protection and data protection for people (Kikerpill, 2020). Another possibility of using Deepfakes is in the political field, to disrupt political campaigns and manipulate public opinion (Chesney & Citron, 2019). Indeed, the Deepfake videos of celebrities may harvest more attention. But more importantly, public figures have a large database of videos, allowing Deepfake to learn and imitate more easily (Dasilva et al., 2021). In the communication realm, fake videos also lead to serious consequences. With the technology of artificial intelligence, the border between true and false videos is further blurred, leading to more difficulty in discerning Deepfake videos. Thus undermining the authority of news and affecting public trust in media (Vaccari & Chadwick, 2020; Godullaet al., 2021).

Currently, the majority of research on the impact of Deepfake focuses on its negative effects, but there are still some researchers affirming the improvements and benefits that Deepfake has brought to certain industries in the applications of virtual reality (Bose & Aarabi, 2019). Liu et al (2019) apply Deepfake's technology to the fashion industry, proposing a deep generative approach called SwapGAN. They use deep learning to simulate a person's fashion style with reference to their pose and body shape. In addition, for the entertainment industry, Deepfake also gives them the opportunity to use more realistic stunt doubles (Godullaet al., 2021).

Detection of Deepfakes

Although most of the Deepfake detection research focuses on technical and algorithmic implementations, there are still some researchers who notice the difficulty for people to identify Deepfake videos.

Khodabakhsh et al. (2019) notice the vulnerability of people when facing fake audiovisual content. By showing participants with fake audiovisual content including Deepfake videos, they find that people rely heavily on the head and facial cues when authenticating. External treatments are provided to track the change in people's Deepfake authentication ability as well. However, the results prove that awareness treatment and financial incentive treatment cannot improve the accuracy and people's ability to detect Deepfake is often lower than estimated (Köbis et al., 2021). In studying the relationship between technical affinity and Deepfake detection, questionnaires are used to examine the relevance of age, education level, and gender on technical affinity, as well as the effect of technical affinity itself on the identification level of Deepfake videos. Ultimately the study concluded that the level of technological affinity possessed by the video viewer was positively correlated with the level of recognition (Kleine, 2022). Some researchers introduced eye gaze tracking technology to collect more precise and accurate data. Combined with questionnaires and eye gaze tracking technology, Tahir et al. (2021) analyze the elements of participants' perceptions of the Deepfake video. In addition, they later developed a customized training program to test the usefulness of some of the video authenticity identification strategies. Finally, the researchers concluded that the participants' perception and identification levels improved significantly after the training. Wöhler et al. (2021) also use eye gaze tracking technology to test participants' reactions to real and deep fake videos. At the same time, participants fill in questionnaires after watching and finally researchers find that the frequency of mouth and eyes is higher.

Audio Deepfake is also a good aspect to specifically examine the perception of Deepfake, many of these studies provide a good reference for Deepfake video research. Focusing on student groups, Watson et al. (2021) examined the dependent variable of students' perception of Audio Deepfake in terms of the audience's knowledge of the technology, audio length, grammatical difficulty, education level, and political factors as independent variables. By examining the probability of identifying Deepfake audio, they concluded that the complexity of the utterance, length, and knowledge of the technology had an impact on the results. Müller et al. (2022) also place their focus on people identifying the authenticity of the Deepfake audio. In this study, participants and the AI simultaneously analyze the authenticity of the unified video to derive differences between humans and computers in recognizing Deepfake videos. The researcher's final point about speech influence on Audio Deepfake detection is meaningful for our project.

Native speakers outperform non-native speakers on audio recognition, while the results on the effect of language on Deepfake video are still missing.

Although most of the previous studies on Deepfake have been focused on Deepfake technology, the key factors and main forgery parts of the four facial manipulation technologies mentioned in the article in the process of Deepfake can still provide a reference for our project in identification strategy. According to Tolosana et al.'s article in 2020, they cite a large number of other people's research results to comprehensively discuss the specific procedures and methods used in each manipulation type in the process of Deepfake and focus on the specific operation of Deepfake technology in facial manipulations, such as Entire Face Synthesis, Identity Swap, Attribute Manipulation and Expression Swap. In addition, Westerlund's article in 2019 summarizes the research and information about Deepfake in existing literature and news reports which claims plenty of imperfections exist in today's Deepfake technology.

Research gaps

Current research about Deepfakes lacks standard measures and overarching theoretical frameworks (Vasist & Krishnan, 2022). Although some strategies and measures have been put forward in some articles on identifying Deepfake videos, these ideas lack practical verification and only have theoretical support. In addition, most of the studies involved the correlation between demographic characteristics, such as the relationship between age and false news acceptance (Guess et al., 2019), while the relationship between Deepfake video perception and demographic characteristics has not been scientifically verified. In general, empirical research is relatively lacking both in the study of the relevant factors of Deepfake video and in the analysis of measures and strategies.

When facing Deepfake videos, people can only rely on their bare eyes and ears, which makes them more vulnerable (Nygren et al., 2021), while current studies of Deepfake detection are more biased from a technological perspective. Lots of researchers are working on improving the algorithm to build machine classifiers for Deepfake videos with better performance (Li et al., 2020; Yang et al., 2019). According to Gurnera et al. (2020), they use Expectation Maximization (EM) algorithm to create a new detection method in the Deepfake videos and receive improved results.

However, with the improvement of Deepfake technology, user detection should be improved as well. Most of the research papers related to Deepfake videos are leaned on the technology part rather than the user detection part. Many researchers noticed the biased situation we mentioned before and made some efforts on helping users to improve their detection ability (Thaw et al., 2020; Köbis et al., 2021). This topic is immature and has great potential for us to investigate. It is crucial to continue working on the users' perspectives. This study will further analyze users' Deepfake video detection with new measures and perspectives.

Methodology

Diary study

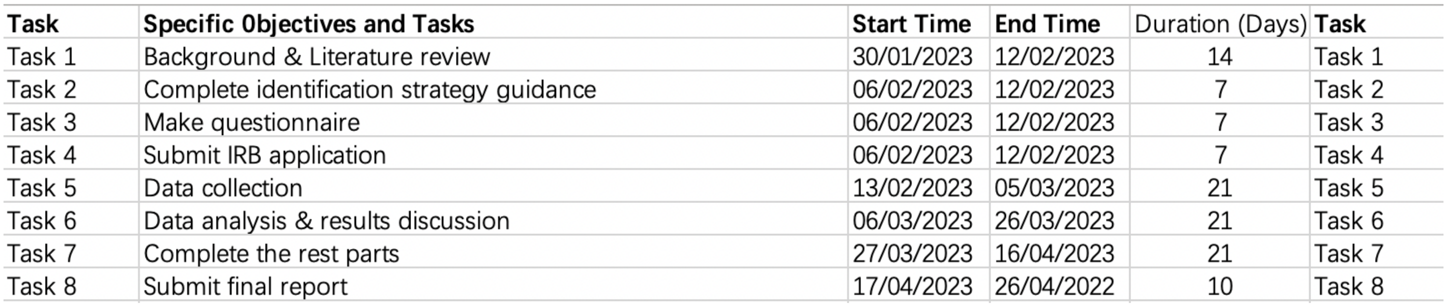
The diary method is a self-reporting research method. Complete and reliable data collection is obtained through the participants' initiative reports (Sun et al., 2011; Bolger et al., 2003). In this project, diary study will be used to track how well people identify Deepfake videos and the factors that influence their perception.

Study design

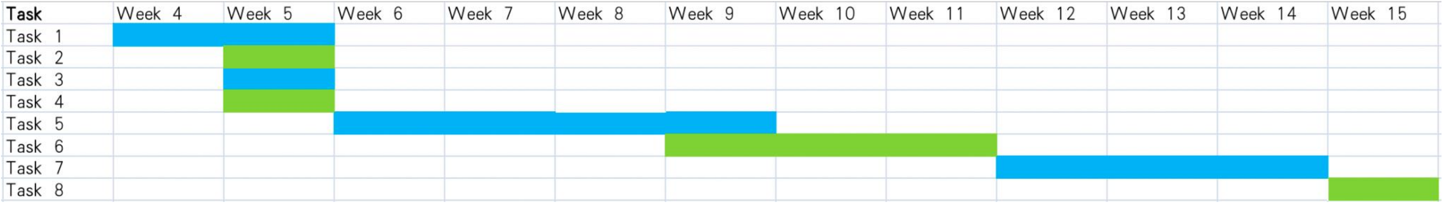
In this diary study, the size of the sample is about 21-24. The duration of the assessment will be 3 weeks and the frequency of the assessment will be once every 7 days.

Firstly, we will provide participants with some instructions about how to respond to the survey and collect their personal data, such as gender, age, academic background, and social media usage. We will also ask them if they have any previous identification experience with Deepfake videos. In addition, we will provide participants with both original videos and Deepfake videos several times in three weeks without letting them know the real aim of the project and ask them to answer the questions we provided in the form of a diary. These include whether they could determine if the video is true or not, and the main reason why they make this decision. In order to avoid the influence caused by the order of the video, we will randomly send those videos to them. We will then repeat those steps above, keep collecting data and see if there are any improvements. Finally, after collecting the participants' diary records, we will interview them about the specific identification process.

Research schedule



Task	Specific Objectives and Tasks	Start Time	End Time	Duration (Days)	Task
Task 1	Background & Literature review	30/01/2023	12/02/2023	14	Task 1
Task 2	Complete identification strategy guidance	06/02/2023	12/02/2023	7	Task 2
Task 3	Make questionnaire	06/02/2023	12/02/2023	7	Task 3
Task 4	Submit IRB application	06/02/2023	12/02/2023	7	Task 4
Task 5	Data collection	13/02/2023	05/03/2023	21	Task 5
Task 6	Data analysis & results discussion	06/03/2023	26/03/2023	21	Task 6
Task 7	Complete the rest parts	27/03/2023	16/04/2023	21	Task 7
Task 8	Submit final report	17/04/2023	26/04/2022	10	Task 8



Task	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15
Task 1	🟦	🟦										
Task 2		🟩										
Task 3		🟦										
Task 4		🟩										
Task 5			🟦	🟦	🟦	🟦						
Task 6						🟩	🟩	🟩				
Task 7									🟦	🟦	🟦	
Task 8												🟩

Dissemination of Results and Contributions of Project

After we finish the research, we will send participants the result of our study, and publish it in the relevant journals.

From users' perspective, this research will probably help participants have a better understanding of those deepfake videos and improve their ability to distinguish deepfake videos when they face some fake news or videos in their daily lives.

From the researchers' perspective, diary studies here will help them establish a different point of view from users' perspectives. They could use similar methods to verify any feasible ways they find to help improve users' ability on deepfake video detection.

Reference

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The state of deepfakes: Landscape, threats, and impact. Amsterdam: Deeptrace, 27.

Albahar, M.A., & Almalki, J. (2019). Deepfake: Threats And Countermeasures Systematic Review.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual review of psychology*, *54*(1), 579-616.

Caldwell, M., Andrews, J.T., Tanay, T., and Griffin, L.D. (2020). *AI-enabled future crime*. Crime Sci. 9,1–13.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, *107*, 1753.

Frenda, S.J., Knowles, E.D., Saletan, W., and Loftus, E.F. (2013). False memories of fabricated political events. *J. Exp. Social Psycho*l. 49, 280–286.

Gamage, D., Ghasiya, P., Bonagiri, V., Whiting, M. E., & Sasahara, K. (2022). Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. *CHI '22: CHI Conference on Human Factors in Computing Systems*, April 2022, Article No.: 103. https://doi.org/10.1145/3491102.3517446

Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with deepfakes–an interdisciplinary examination of the state of research and implications for communication studies Der Umgang mit Deepfakes–Eine interdisziplinäre Untersuchung zum Forschungsstand und Implikationen für die Kommunikationswissenschaft.

Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 666-667).*

Kikerpill. (2020). Choose your stars and studs: the rise of deepfake designer porn. *Porn Studies* (Abingdon, UK), ahead-of-print(ahead-of-print), 1–5. https://doi.org/10.1080/23268743.2020.1765851

Kleine, F. (2022, August). Perception of Deepfake Technology - The Influence of the Recipients' Affinity for Technology on the Perception of Deepfakes. Retrieved January 26, 2023

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021, October 28). Fooled twice: People cannot detect deepfakes but think they can. iScience. Retrieved January 26, 2023

Li, Y., Sun, P., Qi, H., & Lyu, S. (2022). Toward the Creation and Obstruction of DeepFakes. In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, & C. Busch (Eds.), *Handbook of Digital Face Manipulation and Detection* (pp. 71–96). Springer Publishing. https://doi.org/10.1007/978-3-030-87664-7_4

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: a large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216.

Müller, N. M., Pizzi, K., &amp; Williams, J. (2022). Human perception of audio deepfakes. *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. https://doi.org/10.1145/3552466.3556531

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022, August 11). Deep learning for deepfakes creation and detection: A survey. arXiv.org. Retrieved January 26, 2023, from https://arxiv.org/abs/1909.11573

Olan, F., Jayawickrama, U., Arakpogun, E.O. et al. Fake news on Social Media: the Impact on Society. Inf Syst Front (2022). https://doi.org/10.1007/s10796-022-10242-z

Pérez Dasilva, J. Á., Meso Ayerdi, K., & Mendiguren Galdospin, T. (2021). Deepfakes on Twitter: which actors control their spread?.

Sundar, S.S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility (MacArthur Foundation Digital Media and Learning Initiative).

Sun, X., Sharples, S., & Makri, S. (2011). A user-centred mobile diary study approach to understanding serendipity in information research. *Information Research*, *16*(3), 16-3.

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., ... & Zaffar, M. F. (2021, May). Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Thaw, N. N., July, T., Wai, A. N., Goh, D. H. L., & Chua, A. Y. (2020). Is it real? A study on detecting deepfake videos. *Proceedings of the Association for Information Science and Technology*, *57*(1), e366.

Tolosana, Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, *64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

Vaccari, C. & Chadwick, A. (2020). Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media Society*, 6(1), 1–13. https://doi.org/10.1177/2056305120903408

Vasist, P. N., & Krishnan, S. (2022). Deepfakes: an integrative review of the literature and an agenda for future research. *Communications of the Association for Information Systems*, *51*(1), 14.

Watson, G., Khanjani, Z., & Janeja, V. P. (2021). Audio deepfake perceptions in college going populations.https://doi.org/10.48550/arxiv.2112.03351

Westerlund, M. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 9(11): 40-53. http://doi.org/10.22215/timreview/1282

Wöhler, L., Zembaty, M., Castillo, S., & Magnor, M. (2021, May). Towards understanding perceptual differences between genuine and face-swapped videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Yang, X., Li, Y., and Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265.