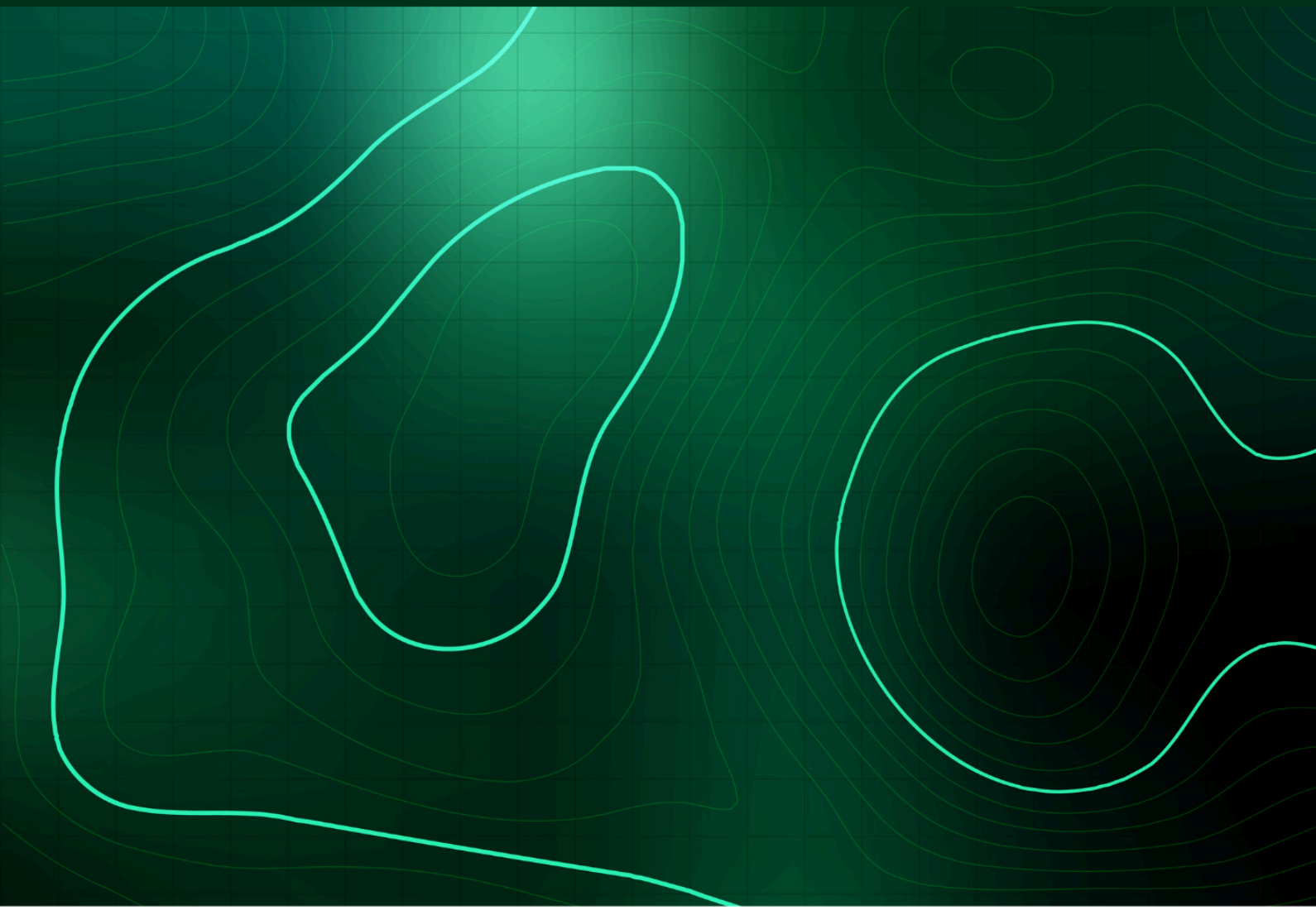
Semester 1 2026

King Abdulaziz University
Faculty of Engineering
General Courses Unit
IE200 - Technical Communication skills

Public Awareness of Deepfake Visuals & Their Potential for Misinformation



Student Name: Mubarak Al-Bishri
University ID: 2537375
Section: 06
Tutor: Owais Hashmi

Abstract

Deepfake technology is capable of ruining personal online reputations, draining bank accounts, and putting societies in conflict, as this technology is getting more advanced and accessible every second. This study aimed to measure the public awareness of deepfake technology, estimate the potential risks of deepfakes in Saudi Arabia, and evaluate the public's ability to detect fake visuals. Data was collected for this study through a survey that consisted of 3 sections, including a practical test and a total of 30 items, distributed online via social media platforms like X, WhatsApp, and Instagram. The participants were a sample of 49 social media users in Saudi Arabia under 70 years old. Data analysis involved generating frequencies and percentages to describe participant awareness levels, detection accuracy, and their perception of deepfakes' potential risks. The findings showed that participants have limited knowledge about deepfake technology. They are more concerned about social risks like reputation damage and privacy invasion than global risks like political misinformation. Also, they are overconfident in their abilities to distinguish fake from real, while the practical test indicated that their performance is poor. This study recommends three solutions: labeling deepfakes, imposing penalties for harmful deepfakes, and highlighting the need for public awareness campaigns to educate the public. Future research could examine how awareness levels and detection ability differ across different groups.

Introduction

Deepfake is a technology that can manipulate or generate all types of media, including videos, images, and audio. Older models' outcomes were visibly imperfect, not so convincing (Rössler et al., 2019). Moving forward in time, nowadays models like the Sora engine 2 are capable of producing highly realistic videos that can convince many people (*Sora*, 2024). In addition to the development in the quality and advancement, it’s getting more accessible, which raises the risks of misapplications by humans. For example, it can be used to spread fake news, misinformation, privacy violations, and damaging reputations (Chapagain et al., 2024; Newswire, 2025). This issue is critical because social media will be heavily affected by it, and this paper argues that most social media users cannot differentiate deepfakes from authentic media due to limited awareness and an underestimation of the technology. The goal of this paper is to use the findings to guide policy development and to help the public be safe from misleading information on the Internet, and to increase their awareness. The objectives of this study include assessing the public’s awareness of deepfake technology, identifying the potential risks associated with deepfake-based misinformation in Saudi Arabia, and evaluating the public’s ability to distinguish between authentic and manipulated visuals.

“The term “deepfake” is derived from the combination of “deep” [referring to deep learning (DL)] and “fake.” It is normally used to refer to the manipulation of existing media (image, video, and/or audio) or the generation of new (synthetic) media using DL-based approaches.” (Altuncu et al., 2024, p. 2). Deepfakes aren't edited or photoshopped videos or images. Deepfakes work through an algorithmic process that alters existing media or creates new content. Deepfakes use two algorithms: the first one is a generator, which creates a training dataset to reach the needed outcome, while the discriminator analyzes how realistic or fake the output is and gives it a grade in different categories. This grade tells the generator what to fix. Both of them work in a loop that eventually leads to a better outcome (*What Is Deepfake Technology?*, 2025). The concept of deepfake was first introduced in the 1990s, but it did not get widespread attention until the 2010s, which is the period when the availability of massive datasets (from the internet), research and development in

machine learning, and more advanced computing resources emerged. All of these factors led to major achievements in the field (*A Brief History of Deepfakes*, 2025).

Recent industry reports also show that this problem is expanding very quickly. For example, the number of deepfake files increased from about 500,000 in 2023 to an expected 8 million in 2025 (Keepnet Labs, 2025). In addition to that, studies show that people now struggle to identify high-quality deepfake videos, and the human accuracy rate can go as low as only 24.5% (Keepnet Labs, 2025). There is also a noticeable growth in voice-based deepfake attacks. Fraud attempts related to synthetic voices increased by more than 1,300% in 2024 (Pindrop, 2025). These statistics indicate that deepfake technology is not only improving in quality, but it is also becoming a serious real-world issue that affects society and public trust in online information.

To answer the objectives mentioned earlier, this research presents several questions that will guarantee the safety of future generations:

1. How much awareness does the public have of deepfakes?
2. What are the potential risks associated with misinformation created through deepfakes in Saudi Arabia?
3. Can the public identify fake visuals from real ones?

Methodology

The primary method of data collection for this project was a survey that included a practical test. This method was chosen because it makes it easier to access most of the population. The purpose of the survey was to collect information about the awareness of the general public and their ability to tell fake from real. The target population for the study was social media users in Saudi Arabia aged under 70 years old. The sample size was 49. The survey was distributed in Arabic and English.

The survey consisted of 3 sections and a total of 30 items. The first section was designed to separate participants into different sections based on what language they prefer. The second and third sections were both in Arabic and English separately, and they are identical. The second section was designed to gather information from participants on their demographics, like age, gender, and nationality. They were included to examine how awareness of deepfakes is affected across the population. In addition to the daily hours they spend on social media, the social media apps they use, prior knowledge about deepfakes, and opinions about deepfakes. The third section (practical test) was designed to examine the participants on their ability to tell. There were two question types: Comparison: Two videos are shown side by side, and the participant chooses which is real or fake. Single video: One video is shown, and the participant decides if it’s real or fake. The survey consisted of 4 types of questions, which are multiple-choice multiple-answer, multiple-choice single-answer, linear scale, and grid scale.

The survey was designed using Google Forms and distributed online. Links to the survey were posted via social media apps like X, WhatsApp, and Instagram. The survey was sent during the period from 2025/10/28 to 2025/11/12.

The data was analysed using frequency and percentage counts in Google Sheets, using functions such as COUNTA, COUNTIF, and COUNTIFS.

Findings & Discussion

The findings presented here are based on the responses to the survey questionnaire distributed to the general public in Saudi Arabia. The main goal of this data is to evaluate public awareness, potential risk, the practical ability to identify deepfakes, and finally, recommended solutions.

RQ1: How much awareness does the public have of deepfakes?

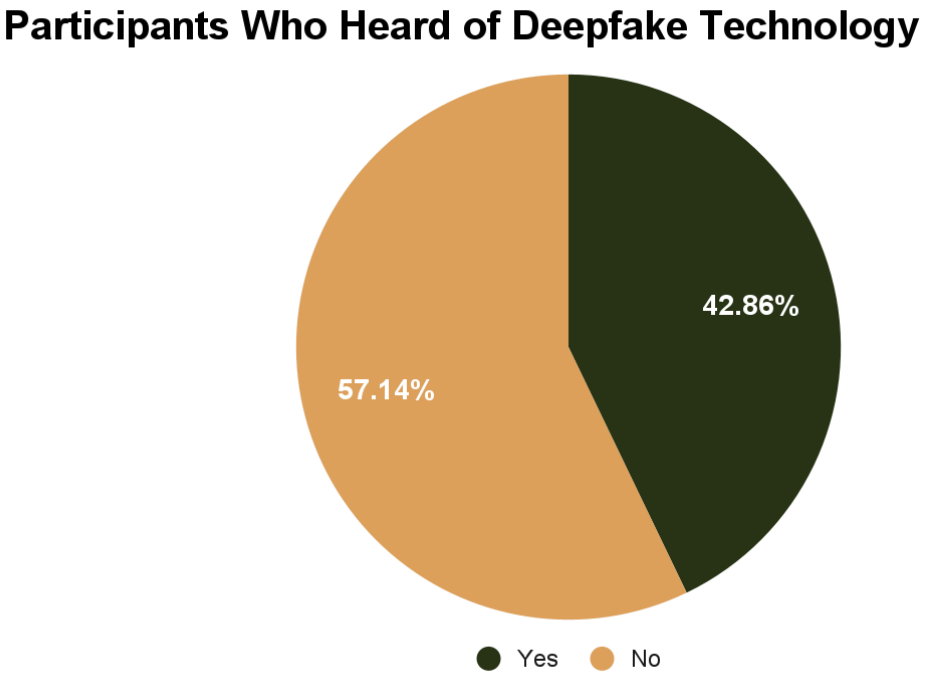


Figure 1: A sample of the general public who heard about deepfakes.

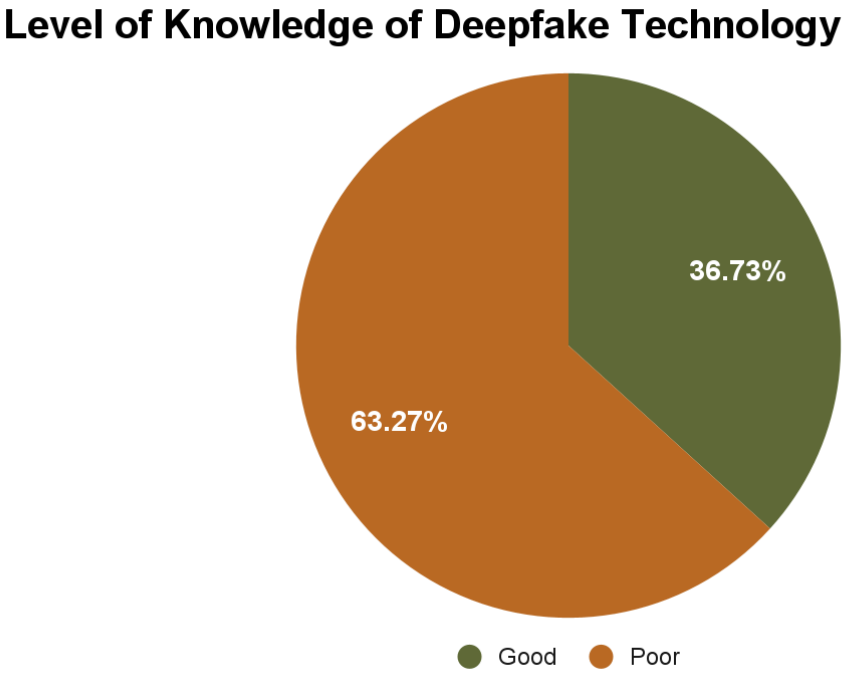


Figure 2: A sample of the general public's level of knowledge about deepfakes.

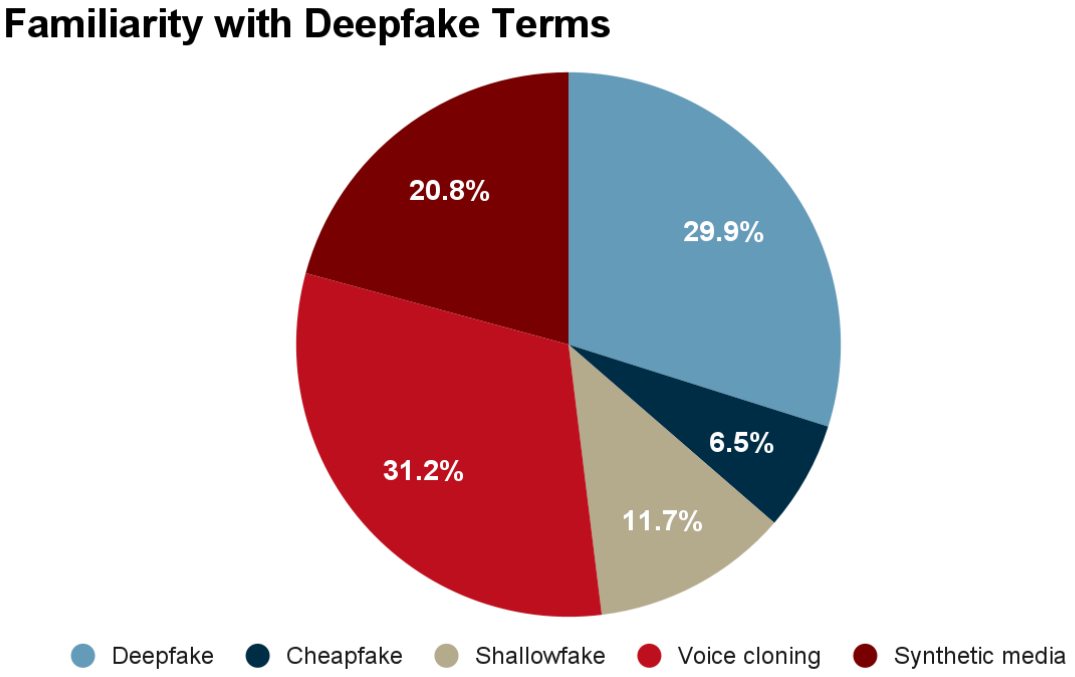


Figure 3: A sample of the general public's familiarity with deepfake terms.

Both Figure 1 and Figure 2 represent that the general public's awareness exists, but they also show that their level of understanding is limited. Figure 1 shows that just under 60% of participants have heard of deepfakes. Figure 2 shows that, among those who have heard of them, only a little over one-third report having a good level of knowledge (ratings 4 and 5). This suggests that most participants have limited knowledge of deepfakes or, in some cases, have not heard of them at all.

Figure 3 illustrates participants' familiarity with specific deepfake terminology. The results indicate that "Voice cloning" was the most recognized term (31.2%), followed closely by "Deepfake" (29.9%) and "Synthetic media" (20.8%). This suggests that participants are more familiar with general uses of deepfake technology, but lack the knowledge of technical terms used in the field, like "Shallowfake" and "Cheapfake".

RQ2: What are the potential risks associated with misinformation created through deepfakes in Saudi Arabia?

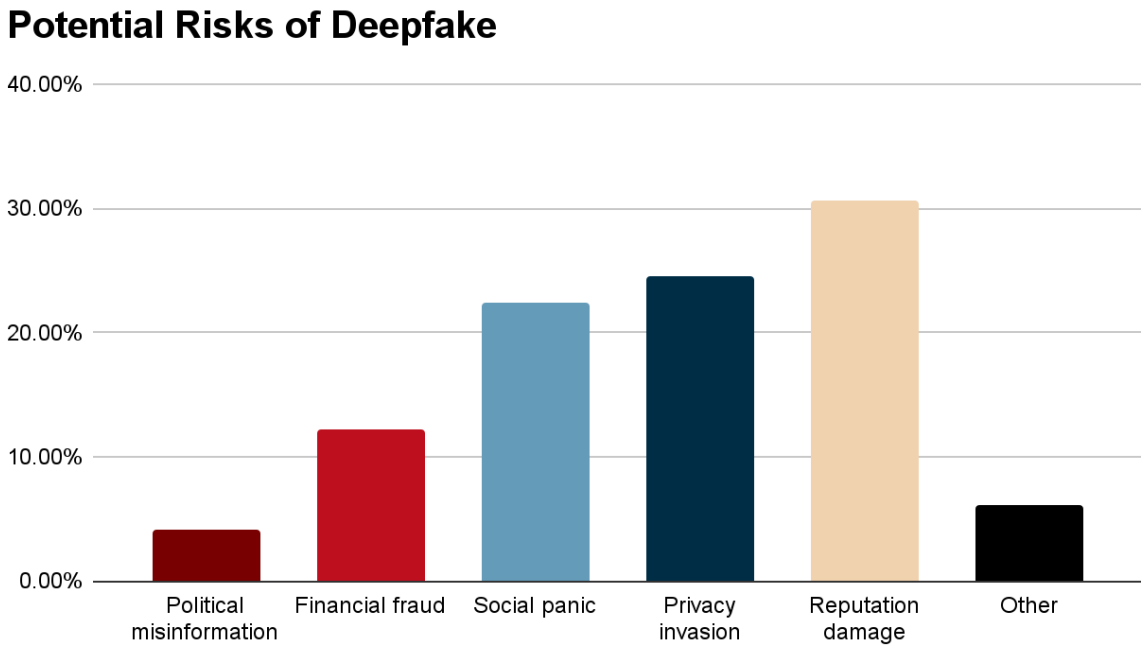


Figure 4: A sample of the general public's awareness of the Potential Risks of Deepfakes.

Figure 4 shows that the participants are already worried about the potential risks of deepfakes. Reputation damage is the most concerning risk, as about 30% of the participants agreed on that. It is followed by privacy violation, where almost a quarter of the participants thought deepfakes can be used to affect their private lives. This agrees with previous research, which also highlighted reputation damage and privacy violations as major ethical concerns (Chapagain et al., 2024). Overall, these data indicate that there are several risks (misapplications) of deepfakes in Saudi Arabia. However, the participants are more worried about social risk (like reputation damage, privacy invasion, and social panic) than worldwide risks like political misinformation.

RQ3: Can the public identify the fake visuals from the real ones?

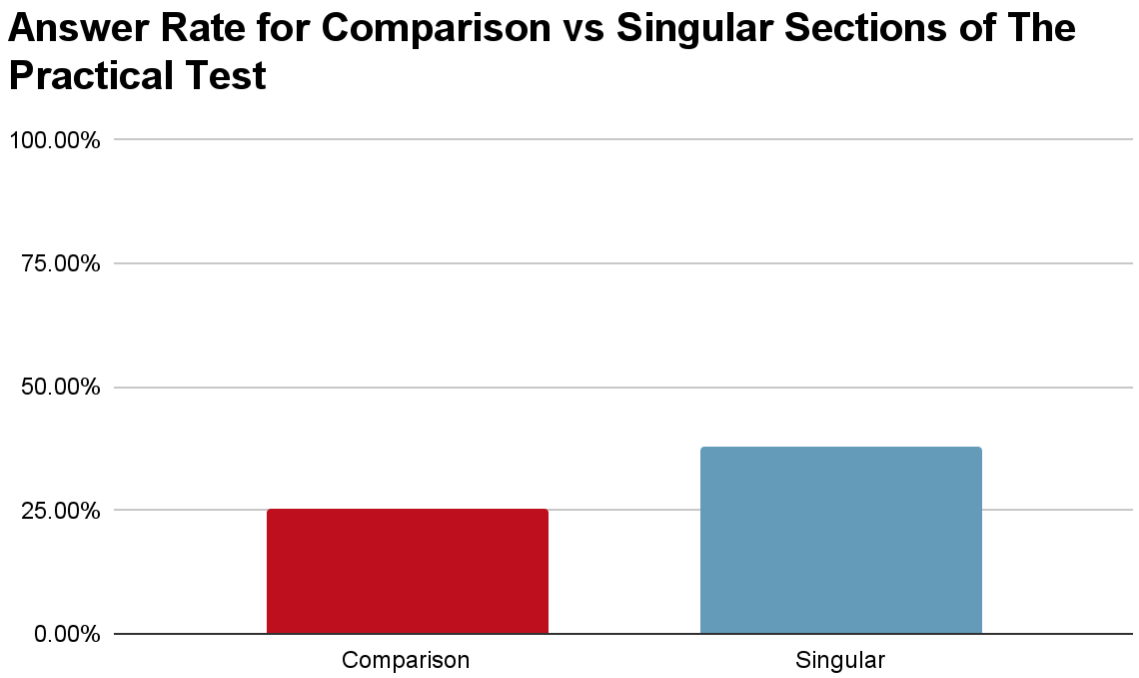


Figure 5: A sample of the general public's ability to tell fake from real by the practical test.

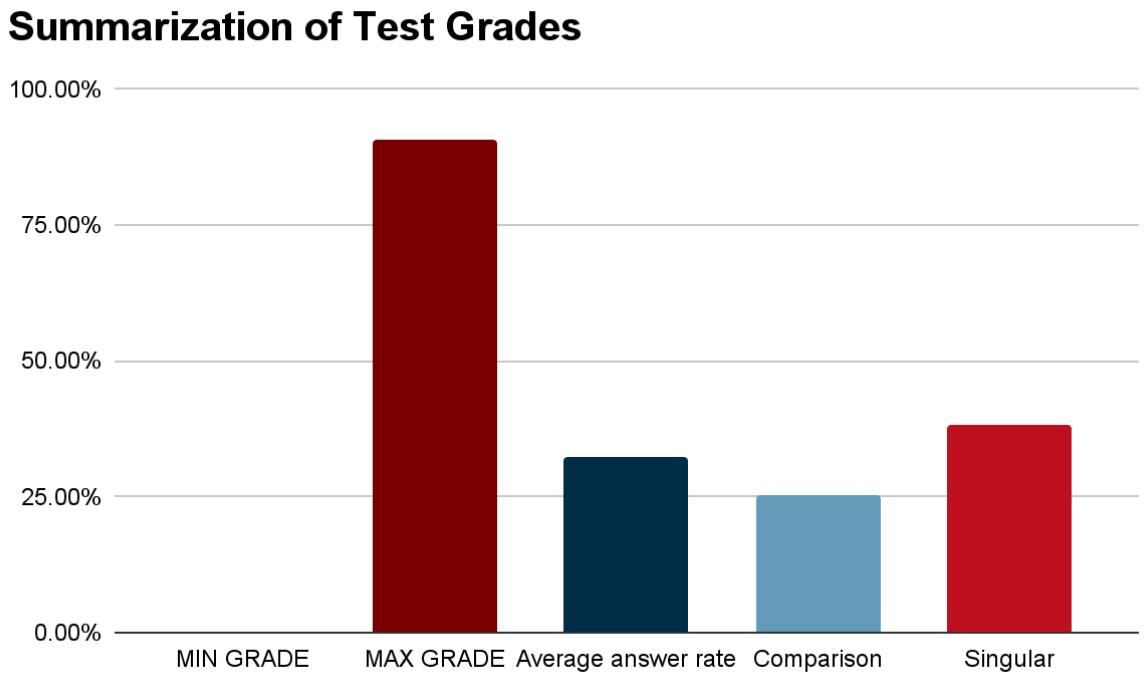


Figure 6: A summary of the practical test.

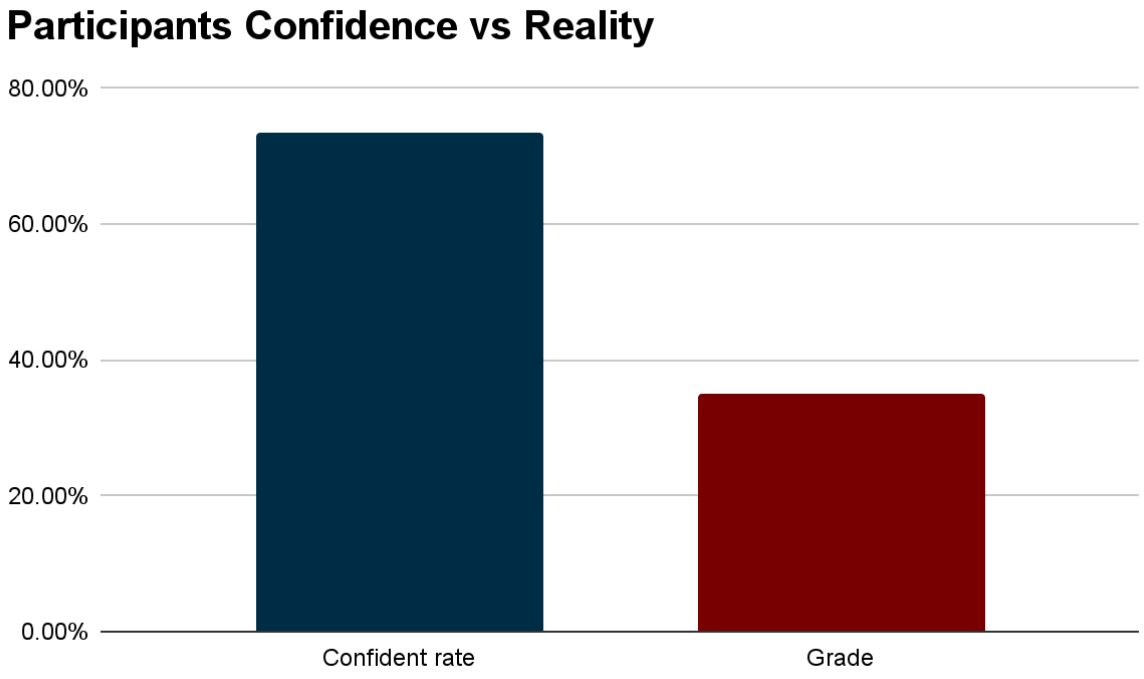


Figure 7: A summary of the confidence rate vs the actual ability to tell

The results of the practical test strongly indicate that the public struggles to distinguish between real and fake visuals. Figure 6 shows that the overall performance was poor, as the average answer rate was only about 35%, while Figure 5 illustrates that participants found it harder to identify fakes in "Comparison" videos compared to "Singular" videos. Figure 7 suggests that the participants are overconfident when compared with their actual performance in the test. It shows that about three-quarters of participants were confident in their ability to tell fake from real, but their actual grade was much lower (35% is the average answer rate). This agrees with previous research, which found that human detection accuracy is often low (24.5%) (Keepnet Labs, 2025). Overall, these data indicate that the public is overconfident, which puts them in a defenseless situation against deepfake misapplications, in addition to the lack of skills to distinguish between real and fake visuals.

Recommendations

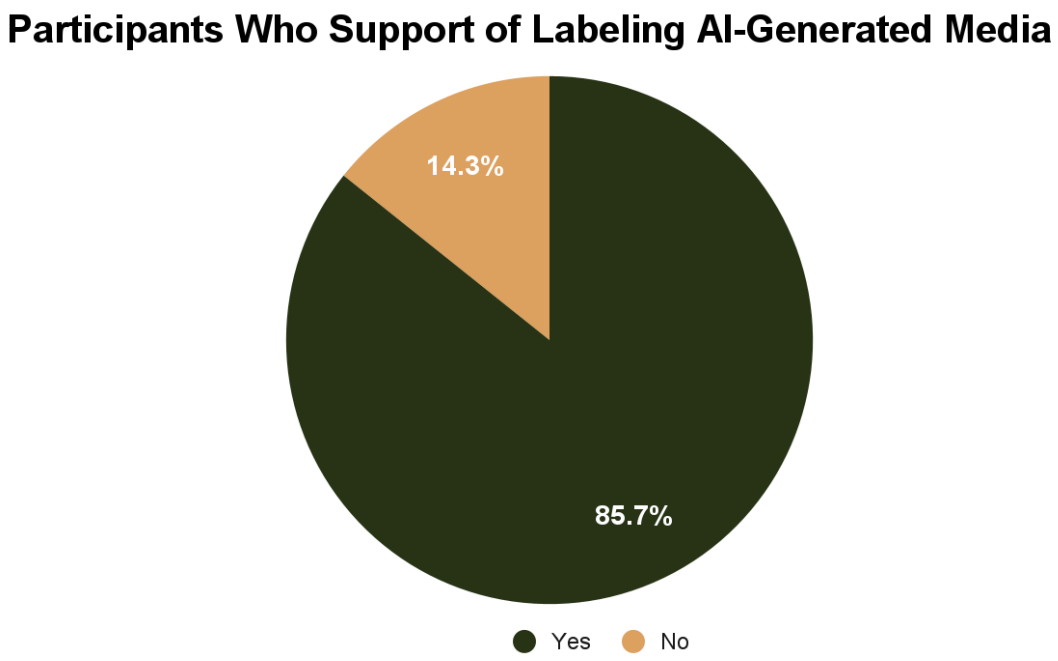


Figure 8: A sample of people's opinions about supporting AI labeling

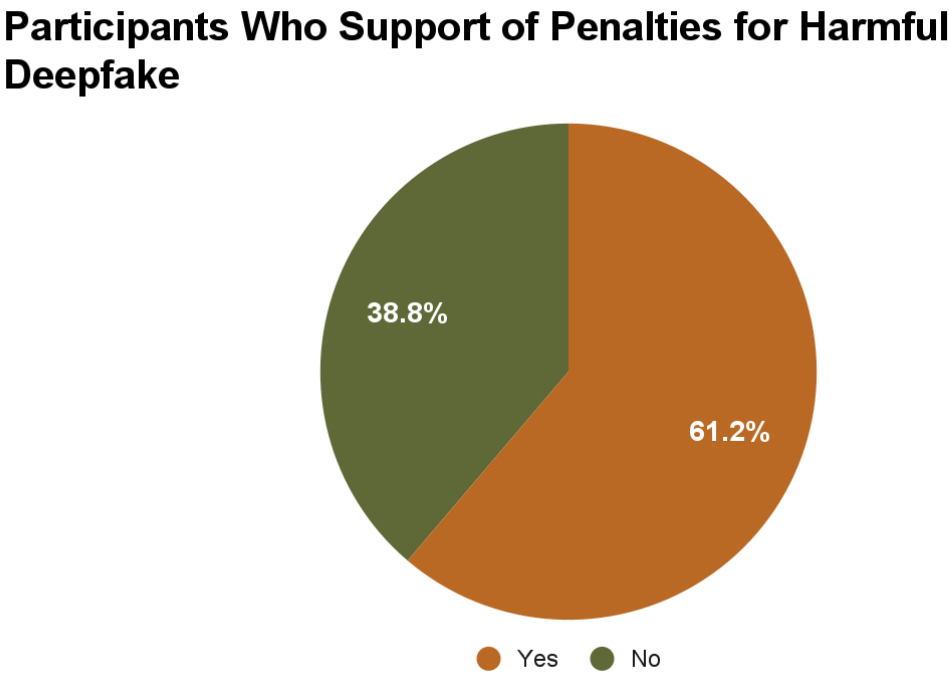


Figure 9: A sample of people's opinions about penalties on harmful deepfakes

After discussing RQ3, it's clear now that the general public of Saudi Arabia lacks the ability to distinguish between fake and real. Both Figures 8 and 9 are optimal solutions for deepfake problems. Figure 8 shows that the majority of the participants support labeling media created by deepfake models (85.7%). This aligns with “technological solutions” and watermarking strategies described by Chapagain et al. (2024), which suggest adding digital marks to media to prove its authenticity. Figure 9 indicates that more than half of the participants want to impose penalties for harmful deepfakes, which goes with the implementation of “regulatory measures” and strict laws to punish those who create criminal content. However, to fully achieve this study's aim, both of the solutions mentioned previously are not enough. The final step to ensure the safety of social media users is to invest in education. Setting up public awareness campaigns that help to raise the level of awareness in the community of Saudi Arabia about deepfake technology is the final step.

Conclusion

In conclusion, this study aimed to measure the public awareness of deepfake technology, estimate the potential risks of deepfakes in Saudi Arabia, and evaluate the public's ability to detect fake visuals. The findings indicate that most participants have limited knowledge of deepfakes. While they do recognize the general uses of the technology, they are unfamiliar with technical terms. The findings also indicate that “reputation damage” and “privacy invasion” were the most concerning risks, to the point that it can be inferred that the participants are more worried about social risk (like reputation damage, privacy invasion, social panic) than global risks like political misinformation. Furthermore, with the most important discovery, the results of the practical test strongly indicate that the public struggles to distinguish between real and fake visuals. Thus, we can say that the public is overconfident, which puts them in a defenseless situation against deepfake misapplications. This study recommends labeling deepfake content through watermarking and imposing penalties for harmful deepfakes. It also highlights the need for public awareness campaigns to raise the level of awareness in the community of Saudi Arabia about deepfake technology. Future research could examine how awareness levels and detection ability differ across different groups.

References

A Brief History of Deepfakes. (2025, July 31). Reality Defender — Enterprise-Grade Deepfake Detection. https://www.realitydefender.com/insights/history-of-deepfakes

Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data*, *7*, 1400024. https://doi.org/10.3389/fdata.2024.1400024

Chapagain, D., Kshetri, N., & Aryal, B. (2024). Deepfake Disasters: A Comprehensive Review of Technology, Ethical Concerns, Countermeasures, and Societal Implications. *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, 1–9. https://doi.org/10.1109/ETNCC63262.2024.10767452

Keepnet Labs. (2025, September 24). *Deepfake Statistics & Trends 2025 | Key Data & Insights—Keepnet*. Keepnet Labs. https://keepnetlabs.com/blog/deepfake-statistics-and-trends

Newswire, I. Q. (2025, October 15). Veo 3.1 vs Sora 2: The Battle of Next-Gen AI Video Generators. *NERDBOT*. https://nerdbot.com/2025/10/15/veo-3-1-vs-sora-2-the-battle-of-next-gen-ai-video-generators/

Pindrop. (2025, June 12). *Pindrop's 2025 Voice Intelligence & Security Report Reveals +1,300% Surge in Deepfake Fraud*. https://www.prnewswire.com/news-releases/pindrops-2025-voice-intelligence--security-report-reveals-1-300-surge-in-deepfake-fraud-302479482.html

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to Detect Manipulated Facial Images* (No. arXiv:1901.08971). arXiv. https://doi.org/10.48550/arXiv.1901.08971

Sora: Creating video from text. (n.d.). Retrieved 3 November 2025, from https://openai.com/index/sora/

What is Deepfake Technology? | Definition from TechTarget. (2025). WhatIs. Retrieved 3 November 2025, from https://www.techtarget.com/whatis/definition/deepfake

Appendices

Appendix A: Copy of the Questionnaire

Choose preferred language / اختر اللغة المفضلة *

- العربي
- English

Nationality *

- Saudi
- Not Saudi

Age *

- 18–24
- 25–34
- 35–44
- 45–54
- 55–70

Gender *

- ○ Male
- ○ Female
- ○ Prefer not to say

How many hours per day do you spend on social media? *

- ○ 1 or less
- ○ 2 - 3 hours
- ○ 4 - 5 hours
- ○ 6 - 7 hours
- ○ More than 7 hours

Which social media platforms do you use at least once a week? *

- [ ] Instagram
- [ ] TikTok
- [ ] X (Twitter)
- [ ] Snapchat
- [ ] YouTube
- [ ] WhatsApp
- [ ] Telegram
- [ ] Other

Have you heard of the term “deepfake” before today? *

- ( ) Yes
- ( ) No
- ( ) Not sure

How familiar are you with the concept of deepfakes? *

	1	2	3	4	5	
Not at all familiar	○	○	○	○	○	Extremely familiar

Which of these terms have you heard before? *

- ○ Deepfake
- ○ Shallowfake
- ○ Cheapfake
- ○ Voice cloning
- ○ Synthetic media
- ○ None

Where did you first learn about deepfakes? *

- ○ Social media
- ○ News or TV
- ○ Friends or family
- ○ School or university
- ○ I haven't heard about it

How often do you see videos or images online that you suspect might be fake or manipulated? *

	1	2	3	4	5	
Never	○	○	○	○	○	Very often

Have you ever shared a video or image that you later found out was fake? *

- ○ Yes
- ○ No
- ○ Not sure

How much do you trust visuals you see on social media to be real? *

	1	2	3	4	5	
Not at all	○	○	○	○	○	Fully trust

To what extent do you agree with the following statements? *
1 = Strongly disagree → 5 = Strongly agree

	1	2	3	4	5
Deepfakes can ...	○	○	○	○	○
Deepfakes can ...	○	○	○	○	○
Deepfakes can ...	○	○	○	○	○
Deepfakes are ...	○	○	○	○	○
Deepfakes can ...	○	○	○	○	○

What do you think is the biggest risk caused by deepfakes? *

- ○ Financial fraud
- ○ Reputation damage
- ○ Political misinformation
- ○ Privacy invasion
- ○ Social panic
- ○ Other

Do you support labeling AI-generated or manipulated videos? *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Do you support legal penalties for harmful deepfakes? *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

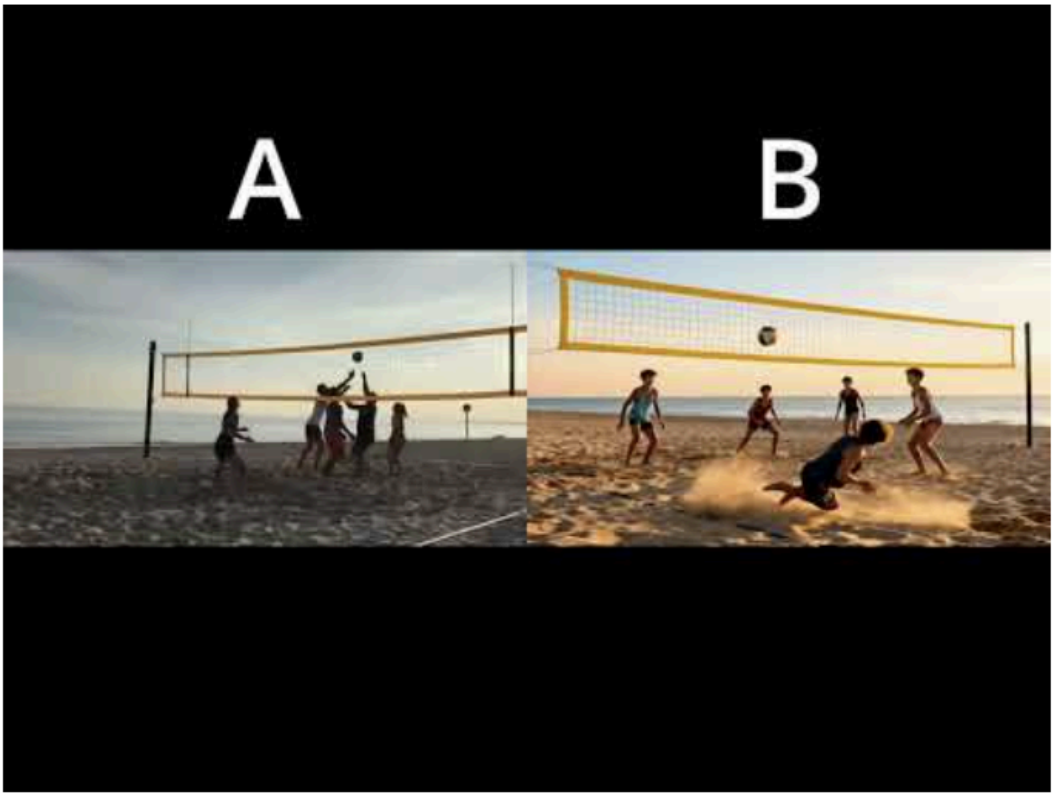
How confident are you in your ability to tell if a video or image is fake? *

	1	2	3	4	5	
Not confident at all	○	○	○	○	○	Very confident

Have you ever identified a fake or deepfake video correctly? *

- ○ Yes
- ○ No
- ○ Not sure

Q1



For the video above, which of the following applies? *

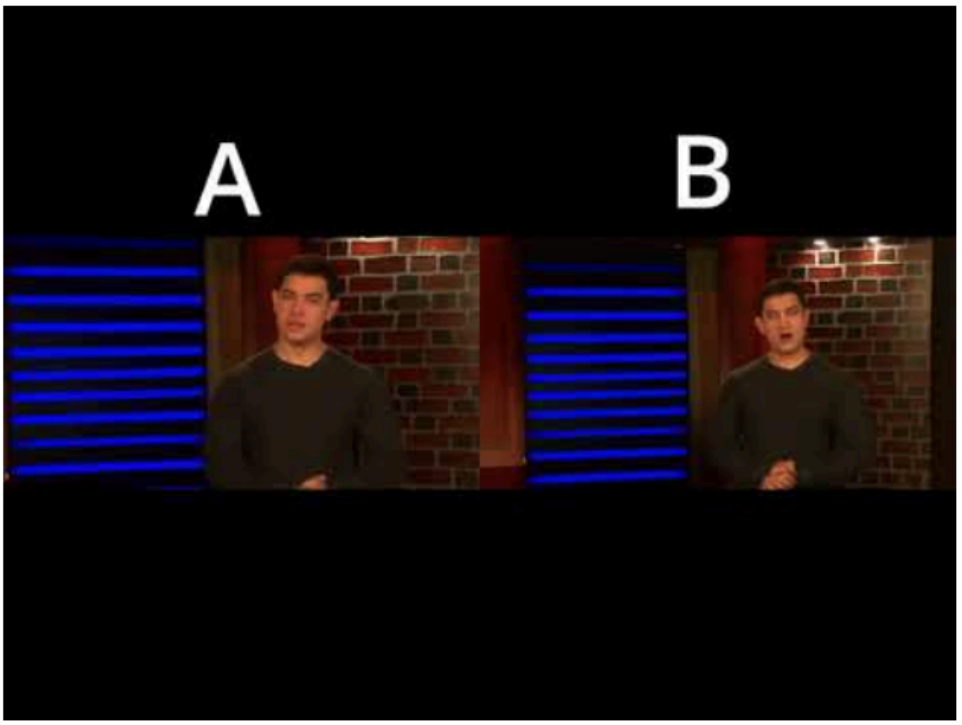
- A is fake
- B is fake
- A & B are fake
- A & B are real
- Not sure

Q2

For the video above, which of the following applies? *

- A is fake
- B is fake
- A & B are fake
- A & B are real
- Not sure

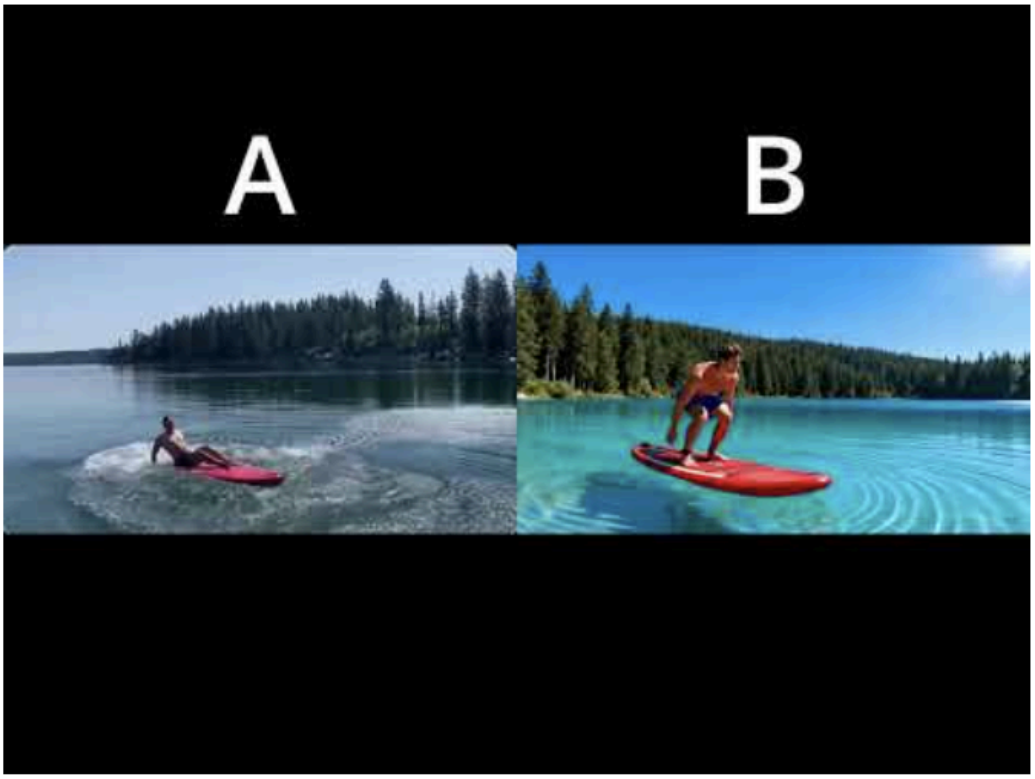
Q3



For the video above, which of the following applies? *

- A is fake
- B is fake
- A & B are fake
- A & B are real
- Not sure

Q4



For the video above, which of the following applies? *

- ○ A is fake
- ○ B is fake
- ○ A & B are fake
- ○ A & B are real
- ○ Not sure

Q5

For the video above, which of the following applies? *

- A is fake
- B is fake
- A & B are fake
- A & B are real
- Not sure

Q6

For the video above, which of the following applies? *

- ○ It's fake
- ○ It's real
- ○ Not sure

Q7

For the video above, which of the following applies? *

- ○ It's fake
- ○ It's real
- ○ Not sure

Q8

For the video above, which of the following applies? *

- ◯ It's fake
- ◯ It's real
- ◯ Not sure

Q9

For the video above, which of the following applies? *

- ( ) It's fake
- ( ) It's real
- ( ) Not sure

Q10

For the video above, which of the following applies? *

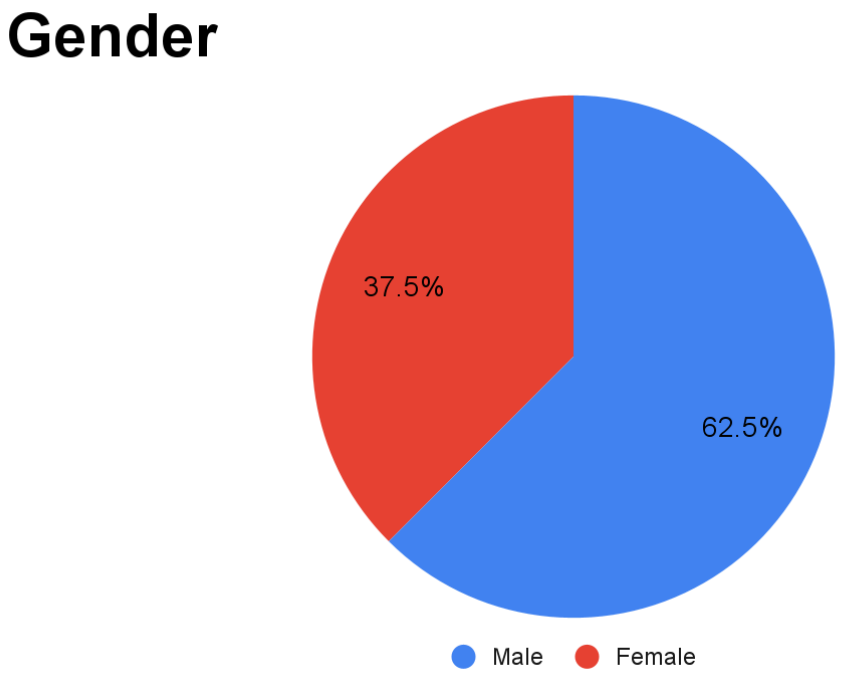
- ◯ It's fake
- ◯ It's real
- ◯ Not sure

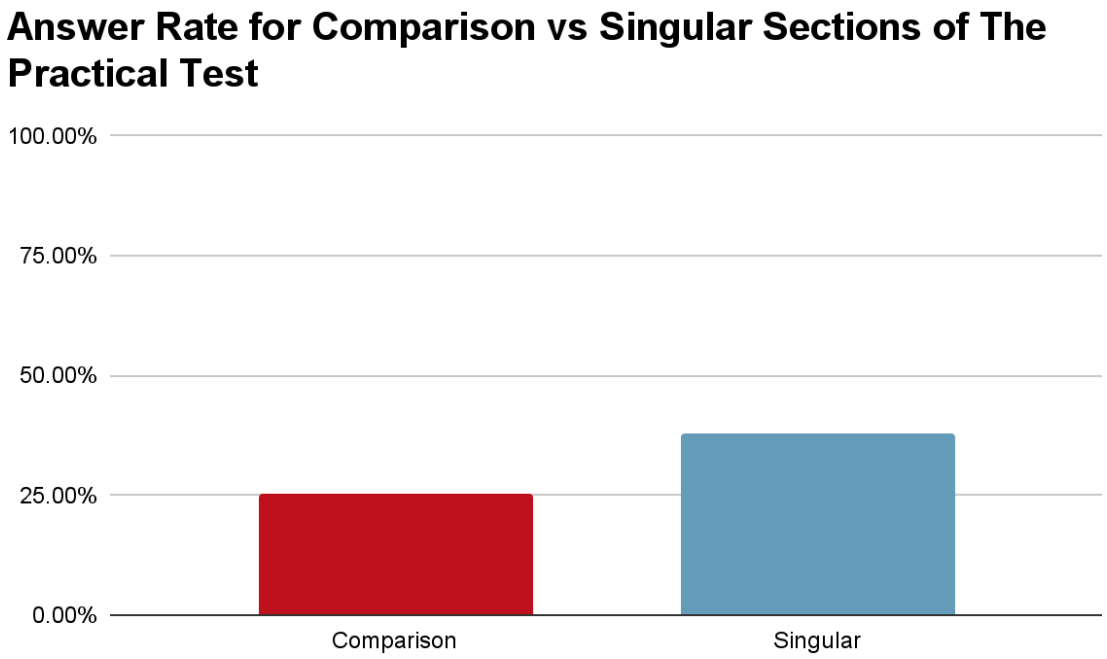
Q11

For the video above, which of the following applies? *

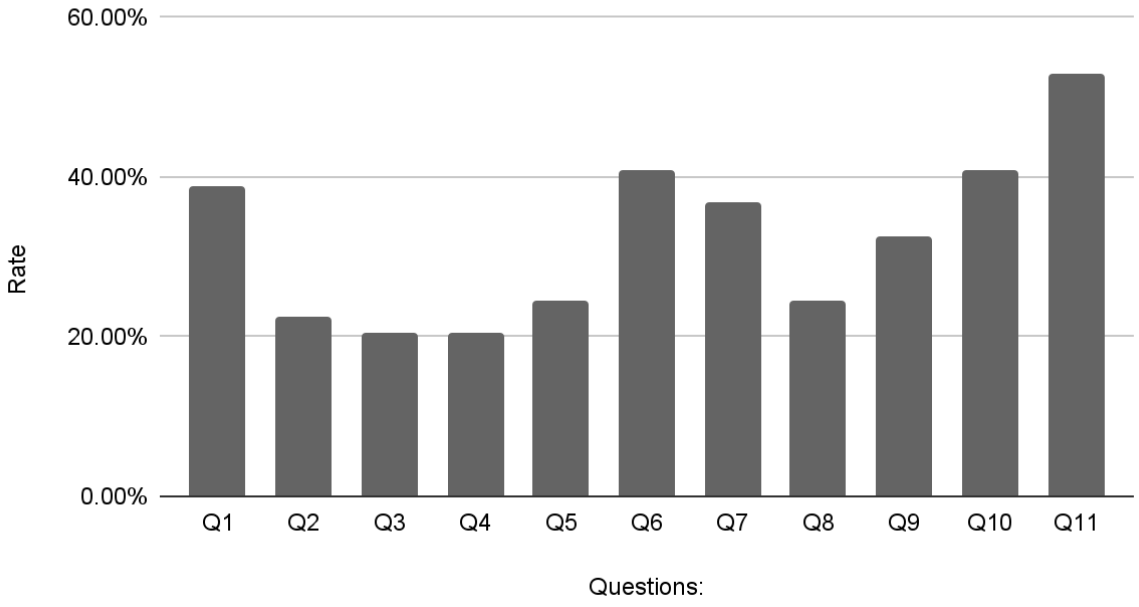
- ○ It's fake
- ○ It's real
- ○ Not sure

Appendix B: Graphs & Charts

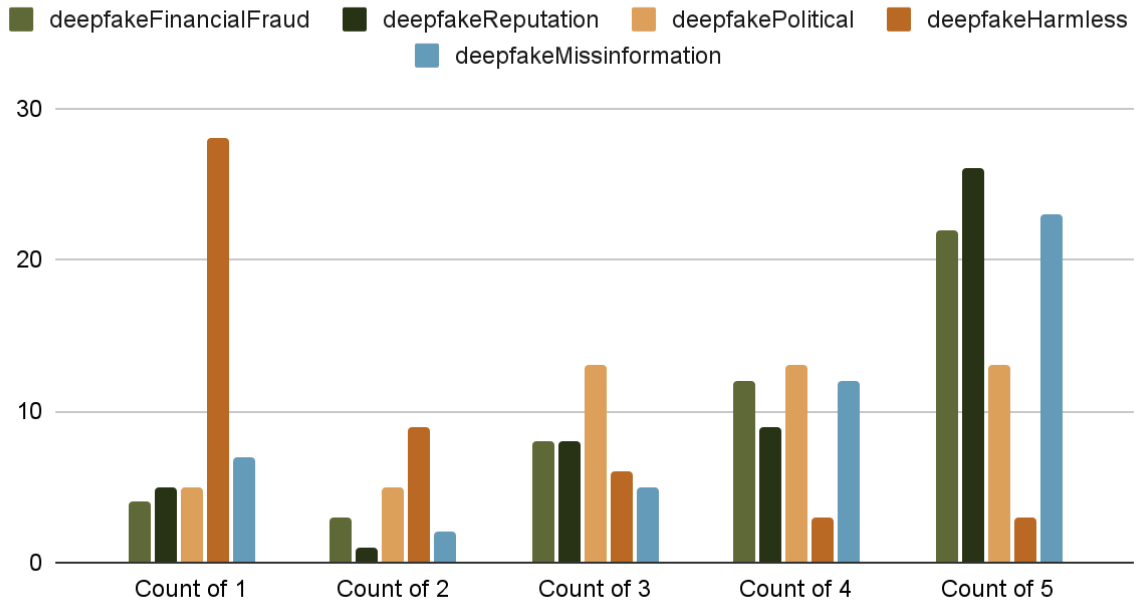




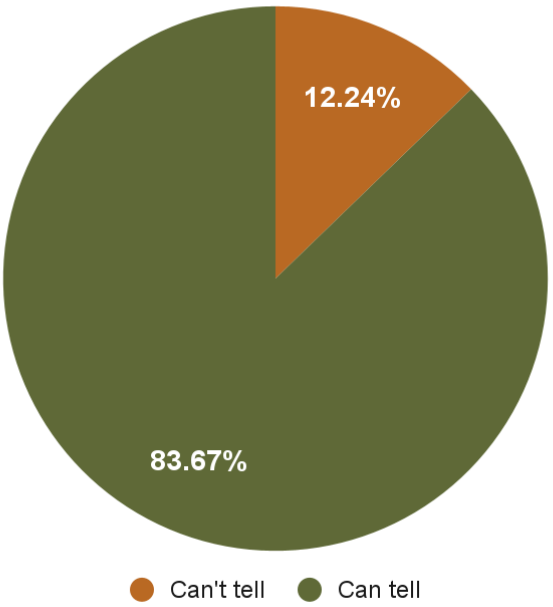
Rate vs. Questions:



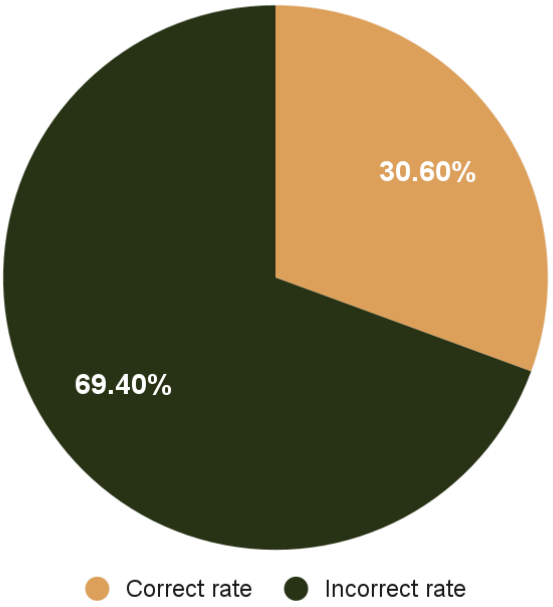
Potintial Risks

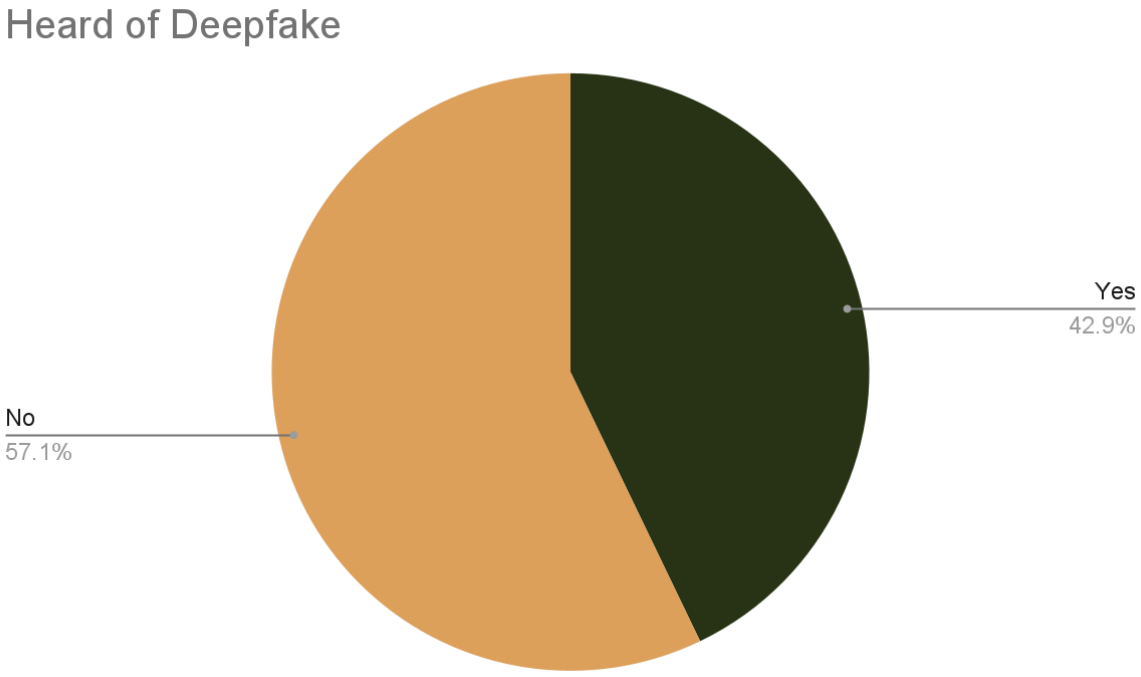


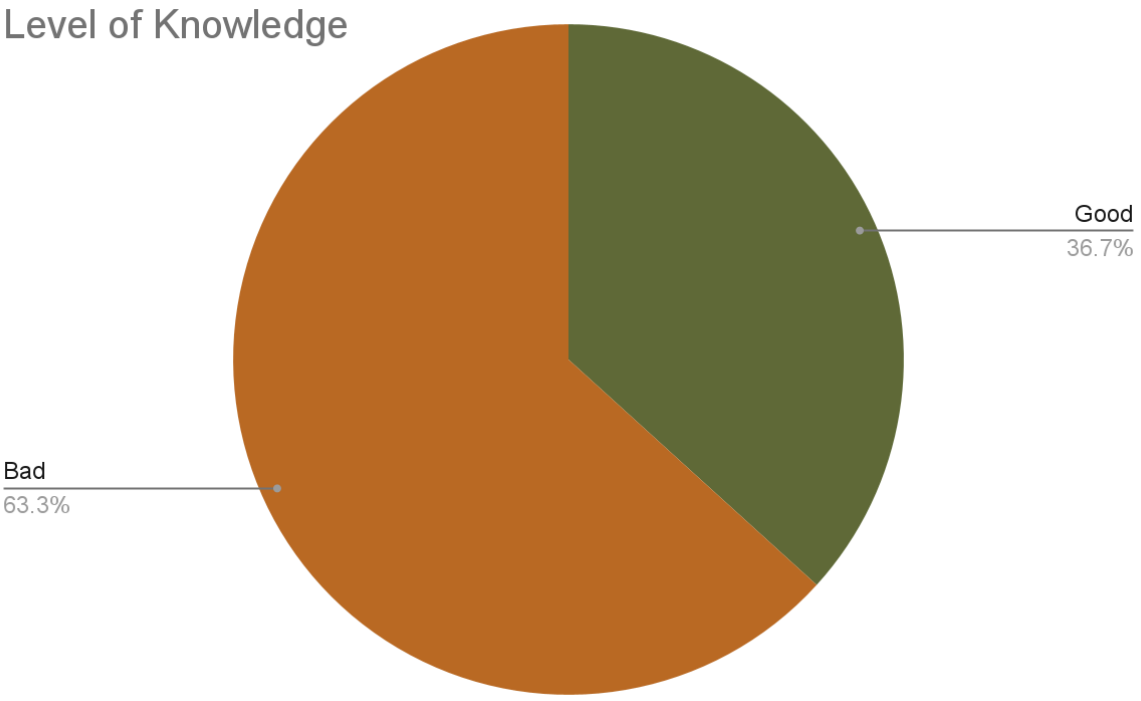
Cofident to Tell Fake from Real

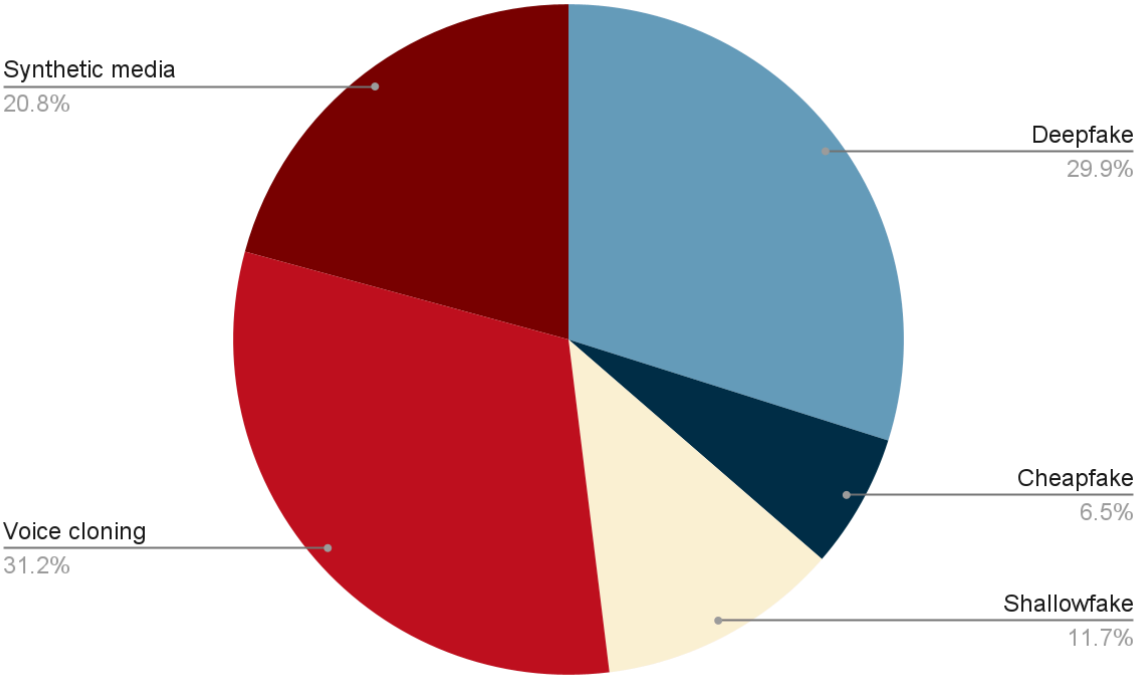


Grading for confidnte people

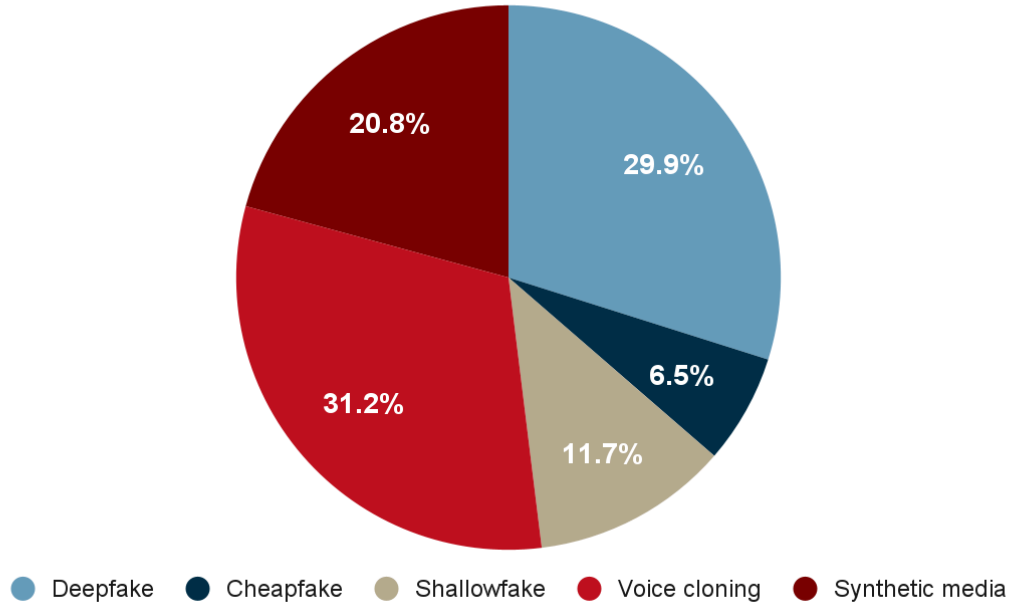



Heard of Deepfake
Yes
42.9%
No
57.1%

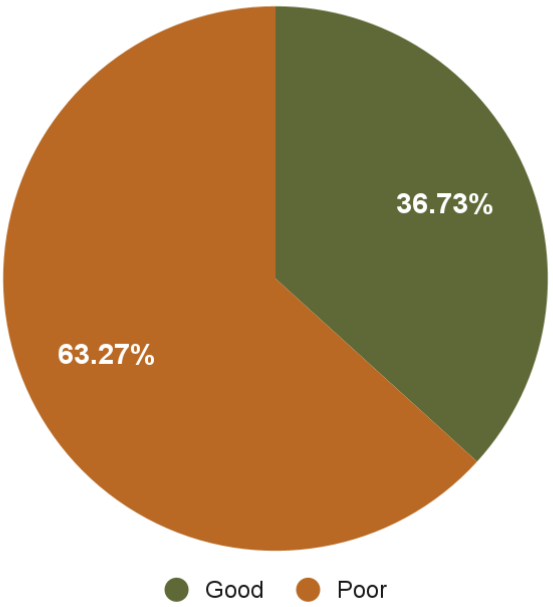

Level of Knowledge
Good
36.7%
Bad
63.3%



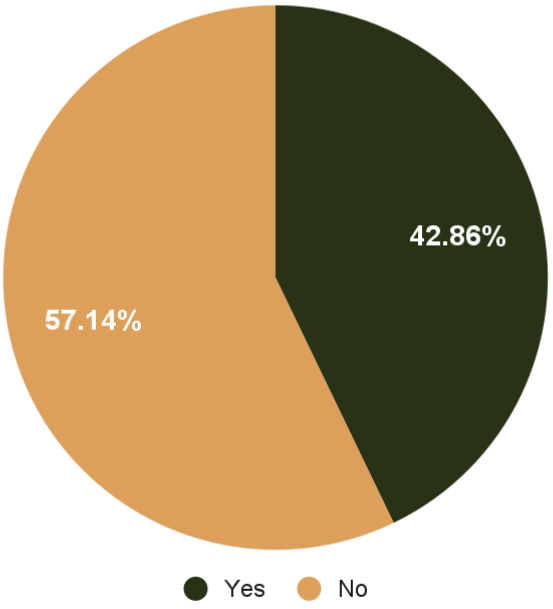
Familiarity with Deepfake Terms



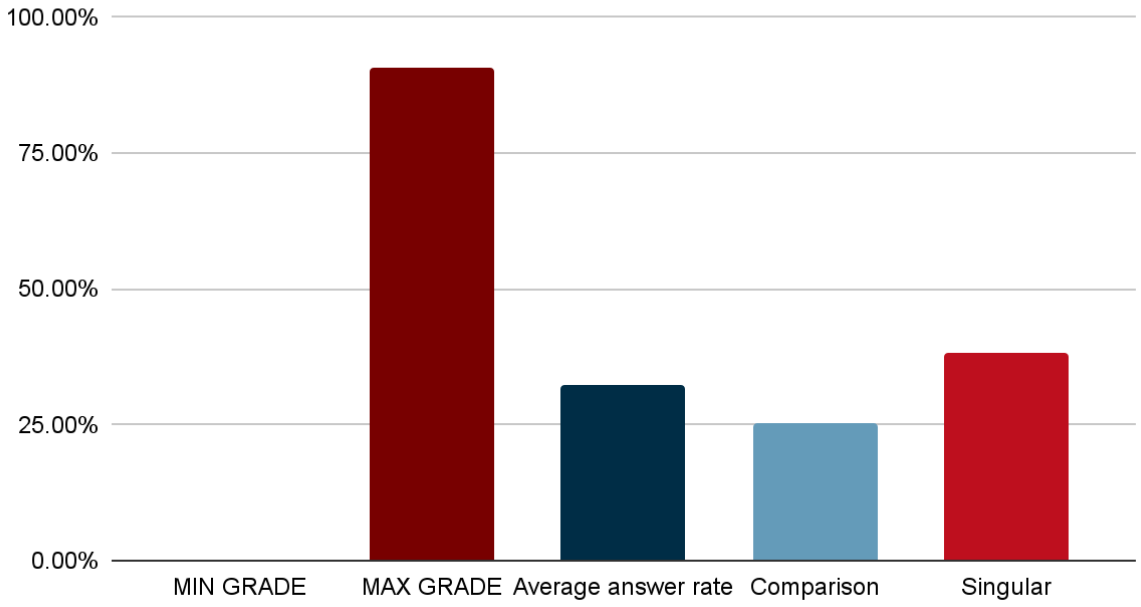
Level of Knowledge of Deepfake Technology



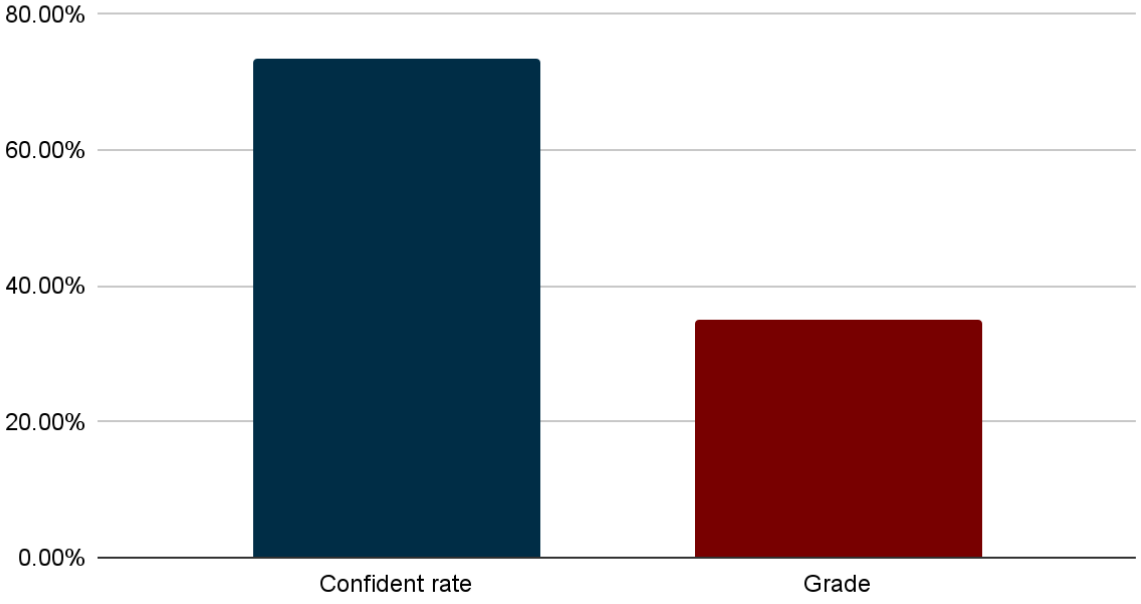
Participants Who Heard of Deepfake Technology



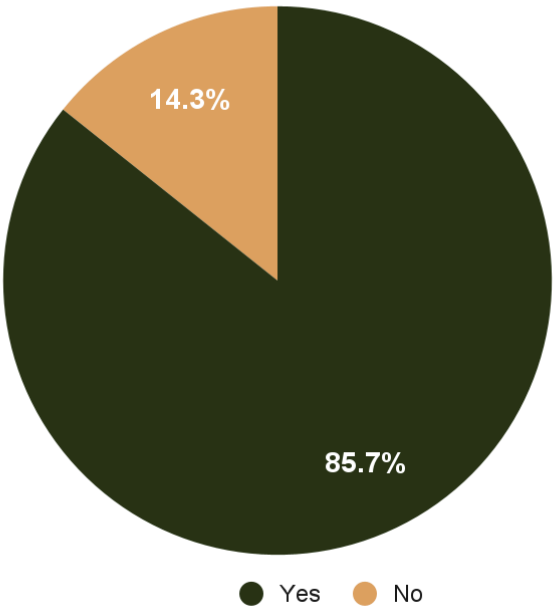
Summarization of Test Grades



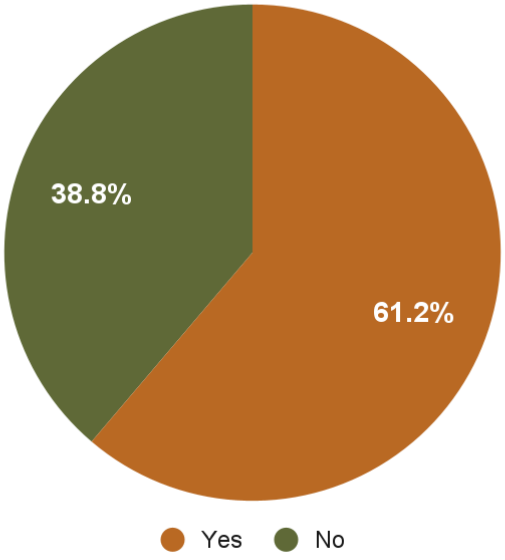
Participants Confidence vs Reality



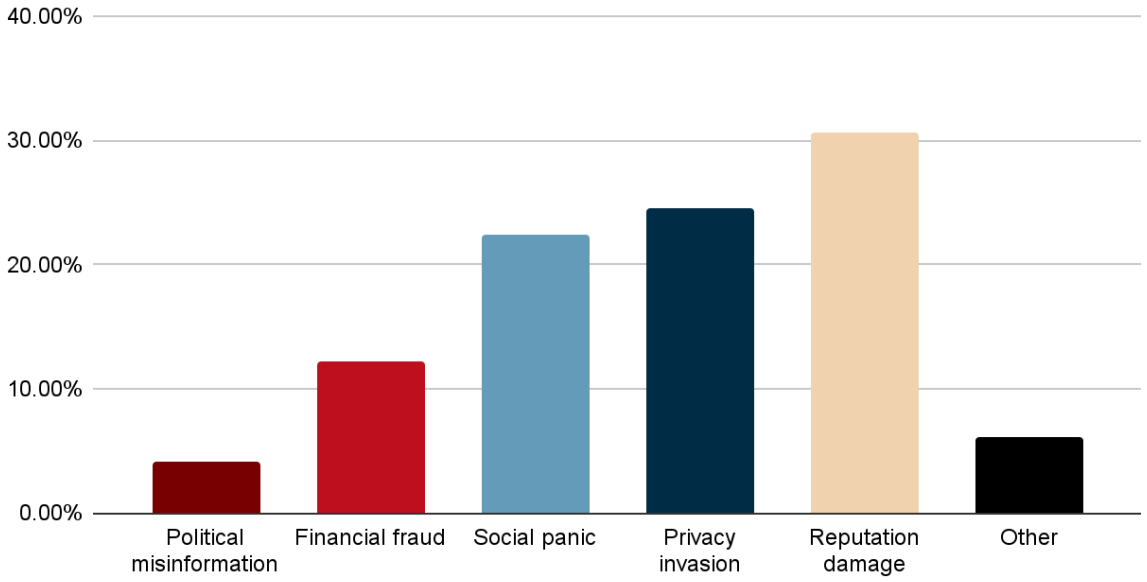
Participants Who Support of Labeling AI-Generated Media



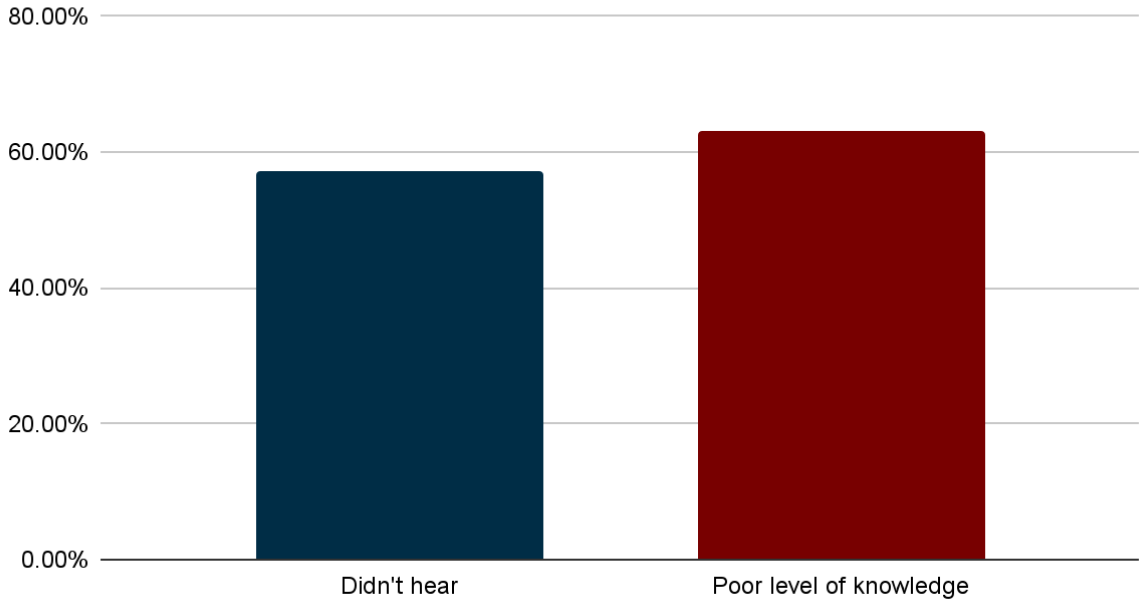
Participants Who Support of Penalties for Harmful Deepfake



Potential Risks of Deepfake



General Level of Knowledge



Familiarity with Deepfake Terms