2024-03-14

IMPROVING HUMAN RECOGNITION OF DEEPFAKES

Jeremy Mumford
Brigham Young University

Honors Thesis

IMPROVING HUMAN RECOGNITION OF DEEPFAKES

by

Jeremy Mumford

Submitted to Brigham Young University in partial fulfillment of graduation requirements for University Honors

Computer Science Department

Brigham Young University

March 2024

Advisor: Quinn Snell

Faculty Reader: Carl Hanson

Honors Coordinator: Seth Holladay

ABSTRACT

IMPROVING HUMAN RECOGNITION OF DEEPFAKES

Jeremy Mumford

Computer Science Department

Bachelor of Science

This thesis is focused on deepfakes, a new term given to fake videos and images generated by deep learning algorithms and models. Deepfakes pose a considerable threat to society by raising the bar for quality in misinformation while also lowering the amount of skill and effort required. Deepfakes threaten to undermine democratic societies by swaying public opinion through misinformation. While many researchers are working hard to develop automated tools to combat deepfakes, this thesis used a 10-item IRB approved survey to examine whether two separate interventions could successfully improve an individual's ability to recognize deepfakes. Demographic differences in recognizing deepfakes was also explored. The results of the survey found that while younger participants responded positively to interventions, older participants reacted adversely to interventions. Older participants also performed significantly worse at recognizing deepfakes.

Acknowledgments

I'd like to thank the Honors Program at Brigham Young University for providing me with the opportunity to develop and execute an undergraduate thesis. Individuals such as Vika Filimoeatu and student employees in the Honors office were essential in keeping me on track.

Thank you, Professor Christophe Giraud-Carrier, for being my initial faculty advisor and inviting me to be a member of his data mining lab. Being part of a research focused group and having a mentor to guide me was critical to my success. While Professor Giraud-Carrier was unable to stay on my committee the entire duration of my project due to retiring after accepting a full-time leadership role in the Church of Jesus Christ of Latter-day Saints, he was an essential part of making this thesis happen. Dr. Quinn Snell stepped in last minute to replace his role.

I also want to acknowledge the other members of my honors committee. Professor Carl Hanson as my reader provided feedback drawn from his research into combating COVID-19 misinformation.

I also want to acknowledge the following technologies that were used to assist me in completing this thesis: ChatGPT Plus (data analysis, crafting and improving content), Prolific (research participant website), Microsoft Word (document writer), Google Colab (Jupyter Notebook IDE + free serverless compute).

Table of Contents

List of Tables and Figures

I. Introduction

Deepfakes are fake images, audio, and videos made using AI tools. When referring to deepfakes in this thesis, the term will be primarily focused on video. These videos are created using deep learning algorithms, specifically generative adversarial networks (GANs), to create realistic but synthetic media content (Goodfellow et al., 2014). The term "deepfake " was coined in 2017 when an anonymous Reddit user named "deepfakes" began sharing manipulated videos of celebrities' faces swapped onto the bodies of adult film actors (Chesney & Citron, 2019). Since then, technology has advanced rapidly, and deepfakes are now used for various purposes, both legitimate and malicious (Maras & Alexandrou, 2019).

There are three types of video deepfakes including head puppetry, face swapping, and lip syncing. Head puppetry involves a complete head and shoulders replacement of an actor by a digital double. Face swapping involves swapping out the face of an existing video. Lip syncing is the final type of deepfakes (Lyu, 2020). For visual examples of deep head puppetry, face swapping, and lip syncing, see Figure 1.

Figure 1: *Deep Fake Examples of Head Puppetry, Face Swapping, and Lip Syncing*

The proliferation of all varieties of deepfakes raises several concerns. First, deepfakes have the potential to erode trust in media and public figures, as it becomes more difficult to distinguish between genuine and fabricated content. The epistemic threat of deepfakes threatens to weaken the legitimacy of all videos viewed on the internet (Fallis, 2020). Second, deepfakes can be used to spread disinformation, which poses a threat to the integrity of democratic processes, such as elections. The 2024 presidential election has seen deepfake audios attempt to dissuade voters from going to the polls (Garrity, 2024). Finally, they can be used to harass, blackmail, and manipulate individuals, leading to significant psychological and social consequences (Citron & Chesney, 2019). There is significant ongoing research on improving automated recognition of deepfakes, causing a game of cat and mouse as perpetrators try to outwit detection. This research and other background will be covered in the literary review. Less research has been done about improving human recognition of deepfakes, an area that this thesis sheds new light on. The discoveries in this thesis highlight that interventions can be effective for younger individuals, and that older generations are most at risk for failing to recognize deepfakes.

II. Literary Background

Human Recognition of Deepfakes. These various concerns brought up by deepfakes highlight the need to examine the human ability to recognize deepfakes. Human ability to detect deepfakes varies and is generally considered to be low. In a study conducted by Nightingale et al. (2020), participants were asked to rate the authenticity of a series of videos, some of which were deepfakes. Results showed that participants were only able to accurately identify deepfakes 48.2% of the time, which is close to chance. This suggests that the average person struggles to discern between genuine and manipulated content. In another study, older individuals were found to be especially vulnerable to deepfakes (Caramancion, 2021). This may be because older people might have less exposure to digital technologies, making them less familiar with the nuances of digitally manipulated content. Additionally, cognitive decline and a potential lack of critical thinking skills may further contribute to the difficulty older individuals face in detecting deepfakes.

Machine Recognition of Deepfakes. In contrast, machine learning algorithms have demonstrated high accuracy in detecting deepfakes. For instance, Afchar et al. (2018) trained a convolutional neural network (CNN) to detect deepfakes, achieving an accuracy of 98.7%. This high accuracy rate underscores the potential of machine learning in combating the spread of manipulated media. Furthermore, Rossler et al. (2019) introduced the DeepFake Detection Challenge (DFDC), a competition designed to encourage researchers to develop advanced deepfake detection techniques. The winning algorithm, submitted by Selvaraju et al. (2020), achieved an accuracy of 82.56%, significantly higher than human recognition rates. This highlights the growing

capabilities of machine learning algorithms in detecting deepfakes, with continued improvement expected as the field advances.

Factors that Affect Machine Recognition of Deepfakes

Technological evolution. Despite the success of machine recognition, there are ongoing challenges. As deepfake technology advances, it becomes harder for algorithms to differentiate between genuine and manipulated content (Thies et al., 2016). This arms race between deepfake creators and detection algorithms can lead to increasingly sophisticated deepfakes that are harder to identify. Additionally, deepfake creators can use adversarial examples—small, carefully designed perturbations to input data—to evade detection by machine learning algorithms (Brown et al., 2017). These adversarial examples exploit vulnerabilities in machine learning models, making it difficult for even the most advanced algorithms to detect deepfakes. This rapid evolution of deepfake technology means that detection algorithms must constantly adapt and improve to keep up. This ongoing challenge requires significant research and development efforts from the machine learning community.

Mismatched expectations. Using an algorithm to spot deepfakes poses unique challenges, as the accuracy is often hard to judge. The good news is that these models can identify what appears to be a perfect deepfake to the human eye. Unfortunately, the inverse is also true. These models can fail to recognize deepfakes with artifacts that are clearly visible to the average viewer (Groh, 2021). This discrepancy highlights the limitations of current deepfake detection algorithms and the importance of continued research to improve their performance. Otherwise, we might be surprised when what seems to be an obvious deepfake slips past the automated guards. As the quality of

deepfakes continues to advance, it is crucial to develop more robust and reliable detection methods that can effectively identify any type of manipulated content.

Law and Ethics Policymakers and technology companies must balance the need for effective detection and mitigation strategies with the protection of privacy and freedom of expression. Who is responsible? Where will detection models be implemented? Who will pay the bill to build infrastructure? Questions like these remain unanswered. Continued collaboration between researchers, policymakers, and technology companies is essential to address this complex issue.

Factors that Affect Human Recognition of Deepfakes

Cognitive Factors. Various cognitive elements are associated with the *human* recognition of deepfakes. Research by Pennycook & Rand (2019) suggests that humans are more likely to believe content that aligns with their existing beliefs, which can make it difficult to recognize deepfakes that confirm those beliefs. This phenomenon, known as confirmation bias, can lead people to accept manipulated content as authentic if it supports their preconceived notions or opinions. Furthermore, individuals with higher cognitive reflection abilities were found to be better at detecting deepfakes (Bronstein et al., 2020). This suggests that critical thinking skills and the ability to question the veracity of presented information can aid in detecting deepfakes. Research by Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. found that people with low levels of what they called 'bullsh*t receptivity' were better at detecting deepfakes. This implies that skepticism and an inclination to question information can be advantageous in identifying manipulated content.

Quality of Deepfakes. Besides cognitive factors and familiarity with deepfakes, the quality of a deepfake can significantly impact human recognition. As technology improves and deepfakes become more realistic, it becomes increasingly difficult for humans to discern between authentic and manipulated content (Thies et al., 2016). High-quality deepfakes may convincingly mimic the appearance, voice, and mannerisms of the subject, making them particularly challenging for human observers to detect. Conversely, poorly executed deepfakes with visible artifacts, such as unnatural facial movements or inconsistent lighting, are more likely to be detected by human observers (Lewis, 2019). These imperfections can serve as clues that the content has been manipulated, allowing vigilant viewers to recognize the deepfake.

Improving Human Recognition of Deepfakes

Education and Awareness. With so many elements conspiring to make deepfakes unrecognizable, increasing awareness of deepfakes and their potential consequences is an essential step toward improving human recognition of manipulated content. By providing people with the necessary knowledge and skills, they will be better equipped to critically assess the content they consume online. Studies have found moderate success in increasing detection by educating people (Iacobucci, 2021). This suggests that targeted educational interventions can have a tangible impact on an individual's ability to recognize deepfakes, helping to build a more informed and resilient digital citizenry.

Tools and Resources. Once the public is educated on the dangers of deepfakes, the next step is to provide them with the tools to combat deepfakes. Developing tools and resources that can aid in the detection of deepfakes can be beneficial for improving human recognition. A great example is Twitter/X's community notes, where members

can flag false content with a warning. Such tools can serve as a valuable resource for individuals who may not have the necessary expertise to identify deepfakes on their own. Providing users with easy-to-use tools can empower them to become more critical consumers of digital media, increasing their confidence in discerning between authentic and manipulated content. By making these resources widely accessible and user-friendly, the public can be better equipped to navigate the digital landscape and mitigate the risks associated with deepfakes.

Collaboration between Human and Machines. Many of these tools empower users to work hand in hand with machines. Combining the strengths of human and machine recognition may be a promising approach to improve deepfake detections. Researchers have found that humans and machine learning algorithms have different strengths and weaknesses when it comes to identifying deepfakes. By combining both together, accurate detection of deepfakes increases significantly (Groh et al, 2021). This collaborative approach underscores the importance of integrating human insight and machine learning capabilities in the ongoing battle against deepfakes. By working together, humans and machines can complement each other's strengths and compensate for their respective weaknesses, resulting in a more robust and reliable deepfake detection system.

III. Methodology

Study Design

This study employed a post-test only randomized controlled trial to determine the effectiveness of two different intervention techniques in helping participants identify deepfake videos. The two interventions included (1) providing participants with written instructions on how to recognize deepfake videos, and (2) providing participants with a video on how to recognize deepfake videos. As this study required human participants, approval was sought and obtained from the Institutional Review Board (IRB). Questions sent to the BYU IRB can reference case # IRB2023-030 for this study.

Research Questions and Hypothesis

The study sought to answer the following research questions:

1. Which intervention techniques best assist individuals in recognizing deepfake videos?

The hypothesis is that any intervention will improve an individual's ability to recognize a deepfake.

2. What are the demographic characteristics of those who recognize deepfake videos versus those who do not?

Existing research has shown that older individuals have a harder time identifying misinformation (Caramancion, 2021).

We'll examine what interactions there are between different demographics and the interventions given to participants. We'll also perform additional analysis on factors such as time taken on the survey and the variation in difficulties between questions to see if there is any notable effect with those variables.

Study Participants

A total of 498 participants were recruited from the research website Prolific, which specializes in providing academic studies with qualified candidates and their anonymized demographic information. Participants were compensated with a small amount of cash ($1.50) upon completion of the survey. Funding was provided courtesy of the Honors Program at Brigham Young University.

Deepfake Dataset

The deepfakes videos used in this survey were pulled from the celeb-df dataset (Li et al., 2019), a large dataset of deepfake videos featuring celebrity interviews. This dataset was chosen out of a wide range of available datasets since it contained actual .mp4 files and high-quality deepfakes. Many other accessible research datasets available in academia are either limited in quality and size, or have the videos saved in sets of images so that machine learning models can process them. The convenience and quality of this dataset made it the best choice for this survey.

Procedures and Data Collection

Once the project on Prolific was activated, 500 participants were redirected to a Qualtrics survey. All participants viewed a consent form and agreed to the terms and conditions. All participants were provided with a short intro that is attached in Appendix 1. The survey software then sorted individuals into one of three intervention groups using a randomization algorithm.

The first group received a control intervention. This intervention provided no information about deepfakes. They were only told that they would be expected to spot manipulated videos. See appendix 1 for the instructions that this group and both other

groups received. Participants in the control group then proceeded immediately to the 10 questions on recognizing deepfakes.

The second group was provided a written explanation of deepfakes and how to recognize them. See Appendix 1 for the written explanation.

The third group was provided a video explaining what deepfakes are and how to recognize them, with visual examples. See Appendix 1 for a link to the video they watched.

Participants then had to go through 10 questions. Each question had the same format. The participant would view a 2 – 10 second silent clip of a celebrity. They would then be asked if they thought the video was manipulated. They had the option of selecting yes or no. Six of the questions were unmanipulated, and four were manipulated. This is to reflect the reality of how most videos we view are unmanipulated. The order of these videos was randomized. The links to the video are in Appendix 2.

At the end of the survey, participants were thanked for their time. They then entered their Prolific id, which allowed their answers to be connected back to the demographic data provided from Prolific. At the conclusion of the study, 498 valid data points were gathered.

Data Analysis

Demographic data on age and gender from the research participant website Prolific was joined together with the survey results of participants. Data was then analyzed using Python code to generate tables and plots. During analysis, rare instances of incomplete demographic information was found. When data on a demographic variable such as age or gender was found missing, that individual's data was excluded

analysis. This is not expected to have a significant impact on the results of the survey. To calculate statistical significance, Python code was used to fit the data to a linear model and apply ANOVA tests using the statsmodel library.

IV. Results

Explanation of Data and Terms

The survey had 10 questions, and each question had a right or wrong answer. This means that a participant could achieve a 'score' of up to 10. The results in this thesis will show mean scores. For example, if a group of two participants scored 6 and 8 out of 10, their mean score would be 7. There are three independent variables of interest in this analysis. The first variable is the intervention group of the participant. This can either be control (no intervention), text, or video. The second variable of interest is gender. The categories are male and female. The third and final variable is age. To balance the need of comparable group sizes and distinct demographics, three age groups were determined. Participants were grouped into age groups of 18 – 29, 30 – 44, and 45+. Below is table 1 showing the group size, mean scores, and standard deviation grouped by independent variable and cross interactions of independent variables. After filtering out bad demographic data, there were a total of 482 participants.

Table 1: *Group size, mean score, and std dev for all categories and cross interactions*

Category	*Intervention*	*Gender*	*Age group*	*# of participants*	*Mean score*	*Std Dev*
Intervention	Control			145	6.63	1.92
Intervention	Text			160	6.55	1.94
Intervention	Video			177	6.58	2.03
Gender		Female		185	6.50	1.97
Gender		Male		297	6.64	1.96
Age Group			18-29	151	6.84	1.87
Age Group			30-44	219	6.74	2.00
Age Group			45+	112	5.96	1.90
Intervention & Gender	Control	Female		71	6.61	2.02
Intervention & Gender	Control	Male		74	6.66	1.82
Intervention & Gender	Text	Female		51	6.27	1.79
Intervention & Gender	Text	Male		109	6.69	2.01
Intervention & Gender	Video	Female		63	6.57	2.07

Intervention & Gender	Video	Male		114	6.59	2.01
Intervention & Age Group	Control		18-29	44	6.43	1.80
Intervention & Age Group	Control		30-44	64	7.13	1.95
Intervention & Age Group	Control		45+	37	6.03	1.83
Intervention & Age Group	Text		18-29	56	7.09	1.79
Intervention & Age Group	Text		30-44	70	6.66	1.96
Intervention & Age Group	Text		45+	34	5.47	1.76
Intervention & Age Group	Video		18-29	51	6.92	1.98
Intervention & Age Group	Video		30-44	85	6.52	2.05
Intervention & Age Group	Video		45+	41	6.29	2.03
Gender & Age Group		Female	18-29	58	6.64	1.86
Gender & Age Group		Female	30-44	72	6.65	2.04
Gender & Age Group		Female	45+	55	6.16	1.00
Gender & Age Group		Male	18-29	93	6.97	1.87
Gender & Age Group		Male	30-44	147	6.78	1.99
Gender & Age Group		Male	45+	57	5.75	1.80
Gender, Age Group & Intervention	Control	Female	18-29	23	6.70	1.89
Gender, Age Group & Intervention	Text	Female	18-29	17	6.76	1.68
Gender, Age Group & Intervention	Video	Female	18-29	18	6.44	2.06
Gender, Age Group & Intervention	Control	Female	30-44	26	6.85	2.17
Gender, Age Group & Intervention	Text	Female	30-44	19	6.42	2.04
Gender, Age Group & Intervention	Video	Female	30-44	27	6.63	1.96
Gender, Age Group & Intervention	Control	Female	45+	22	6.23	2.02
Gender, Age Group & Intervention	Text	Female	45+	15	5.53	1.41
Gender, Age Group & Intervention	Video	Female	45+	18	6.61	2.33
Gender, Age Group & Intervention	Control	Male	18-29	21	6.14	1.68
Gender, Age Group & Intervention	Text	Male	18-29	39	7.23	1.84
Gender, Age Group & Intervention	Video	Male	18-29	33	7.18	1.91
Gender, Age Group & Intervention	Control	Male	30-44	38	7.32	1.79
Gender, Age Group & Intervention	Text	Male	30-44	51	6.75	1.95

Gender, Age Group & Intervention	Video	Male	30-44	58	6.47	2.10
Gender, Age Group & Intervention	Control	Male	45+	15	5.73	1.53
Gender, Age Group & Intervention	Text	Male	45+	19	5.42	2.04
Gender, Age Group & Intervention	Video	Male	45+	23	6.04	1.77

Analysis of Interventions

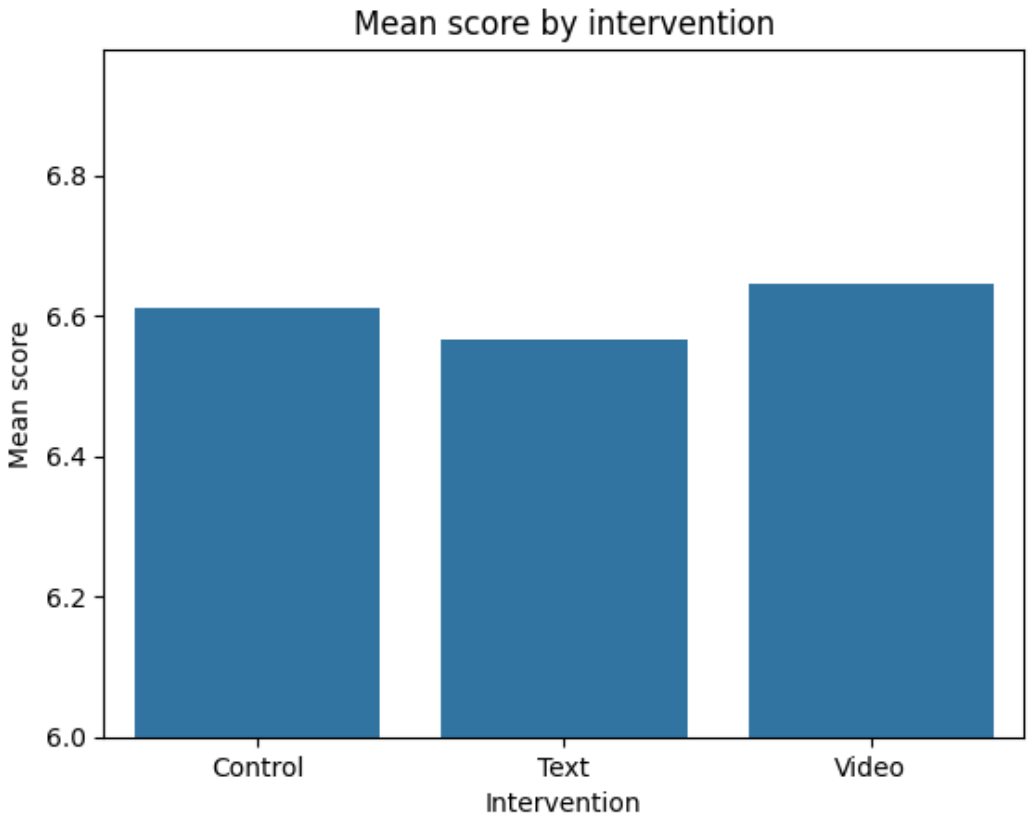

Figure 2: *Mean score by intervention*

Initial analysis of the intervention methods would indicate little effect. The control group scored a mean of 6.6 points. The text and video intervention mean scores are practically identical, with a p-value of .93. However, this is not the full story.

Analysis of Demographics

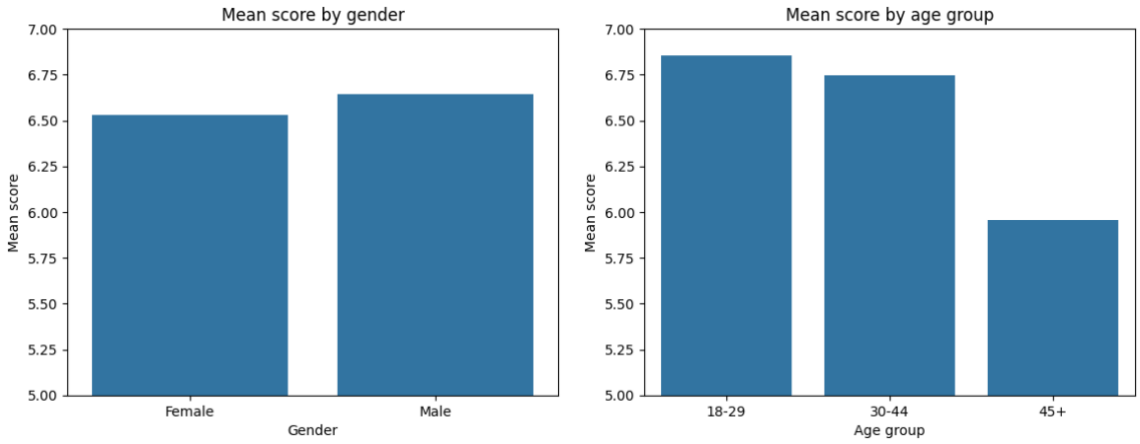

Figure 3a: *Mean score by gender*, Figure 3b: *Mean score by age group*

The difference in scores between male and female are minimal. Practically no difference exists. Running a t-test gives a p-value of 0.44 confirming that we cannot reject the null hypothesis of no difference in how male and female participants perform on identifying deepfakes videos. However, the difference between age groups is much larger. Running an ANOVA test on age groups shows a statistically significant difference ($p = 0.0004$) in the mean scores by age. Older participants show a significant downward trend in their performance. 18 – 29 and 30 – 44 had a mean score of 6.8 and 6.7 respectively, while the older group age 45 and older saw a mean score drop to 5.9 out of 10. An alternate way of viewing this is that 18 – 29 had a mean score of 68%, while 45+ had a mean score of 59%. We can reject the null hypothesis that there is no difference in the ability of different age groups to identify deepfake videos.

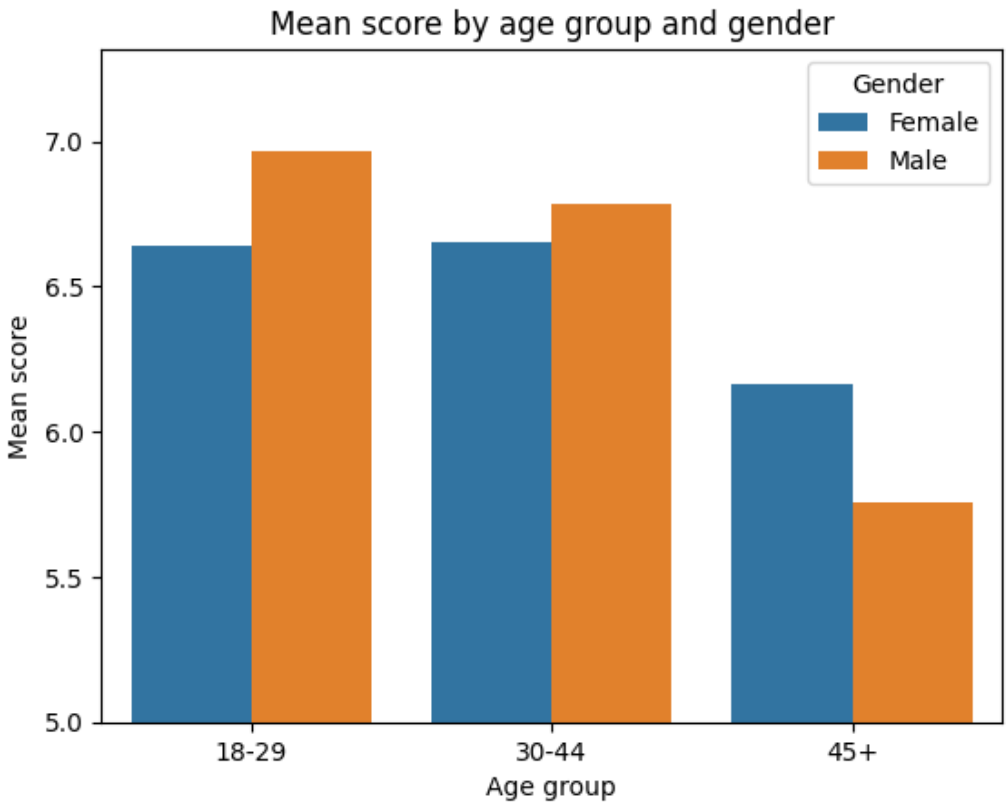

Figure 4: *Mean score by age group and gender*

Breaking the scores down by both gender and age demographics, unusual differences start to emerge. The direction of the differences in gender changes between different age groups. While this difference is not statistically significant ($p = 0.30$), it hints that the interactions between different variables will reveal unexpected results. The next section of analysis continues this trend.

Analysis of Interventions within Demographics

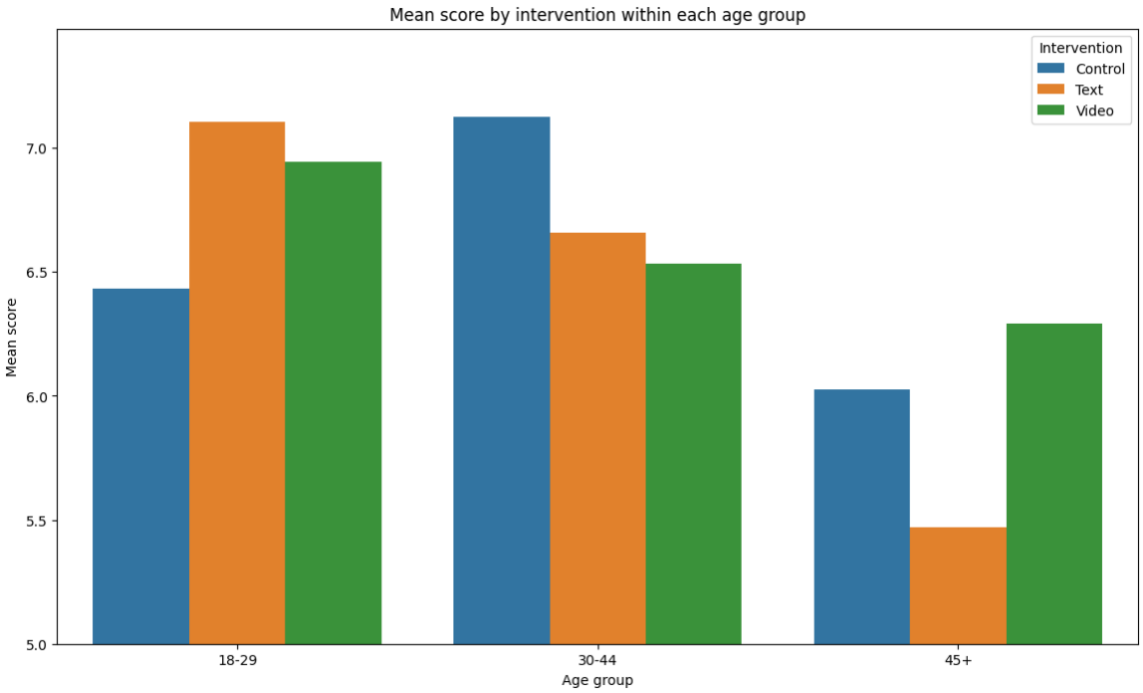

Figure 5: *Mean score by intervention within each age group*

Looking at the cross between age and interventions, the differences become more apparent. Younger participants saw the greatest improvement from interventions, with a half point improvement. The age group of participants ages 30 – 44 had a mean score exceeding 7 out of 10 questions correct in the control group, but their mean score dropped by half a point with either type of intervention. Older participants did the worst in the control group with a mean score of 6 in the control group. Surprisingly, they did somewhat better with video but dropped half a point on average with a text intervention. Running an ANOVA test on this cross-interaction finds that this result is statistically significantly ($p = 0.04$). The trends between different age groups effectively cancel each other out on the aggregate level, which is why the base analysis revealed practically no difference.

What about the differences between male and female? Can a similar interaction be found where differences cancel out on the aggregate level?

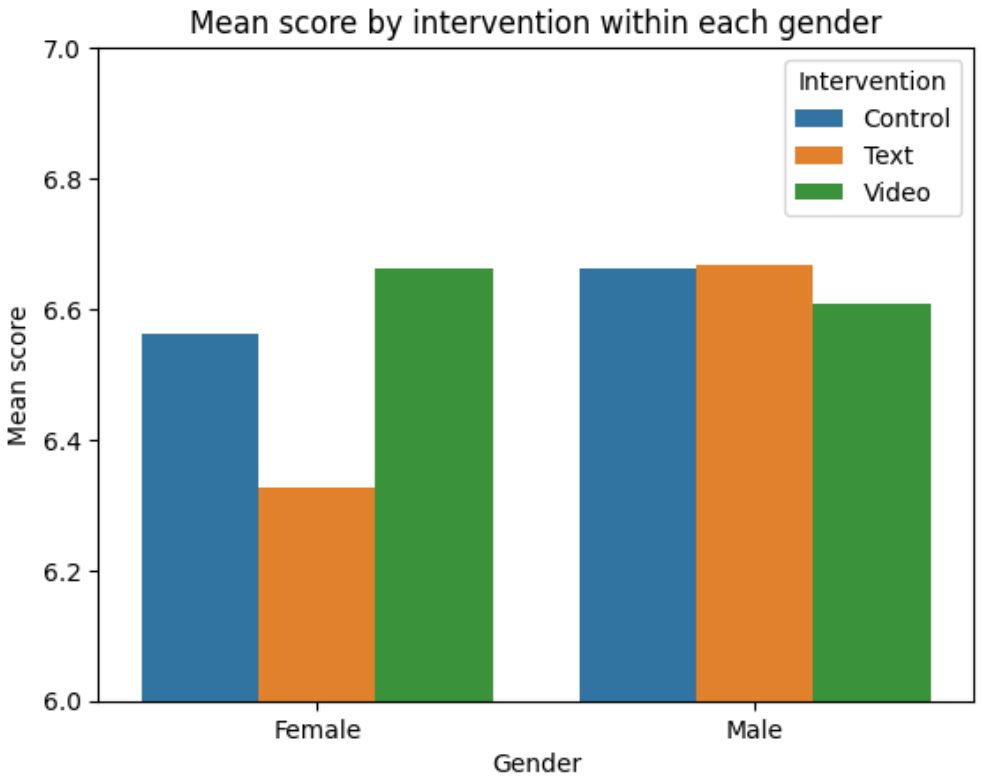

Figure 6: *Mean score by intervention within each gender*

The differences between male and female are less pronounced than the differences seen with age groups. For men in particular, the difference is almost non-exist. The p-value from ANOVA tests is 0.64, showing that this cross-interaction is not statistically significant. However, the earlier analysis of demographics showed that the differences between male and female varied between different age groups. This indicates that all three variables should be looked at together.

Combining an analysis of all three variables reveals these stronger differences. For ease of viewing, two graphs have been made to show how filtering on male and female affects the mean scores (see figures 7 and 8). Note that showing the cross

interaction of three variables reduces most of these group to around 20 or 30 participants in size.

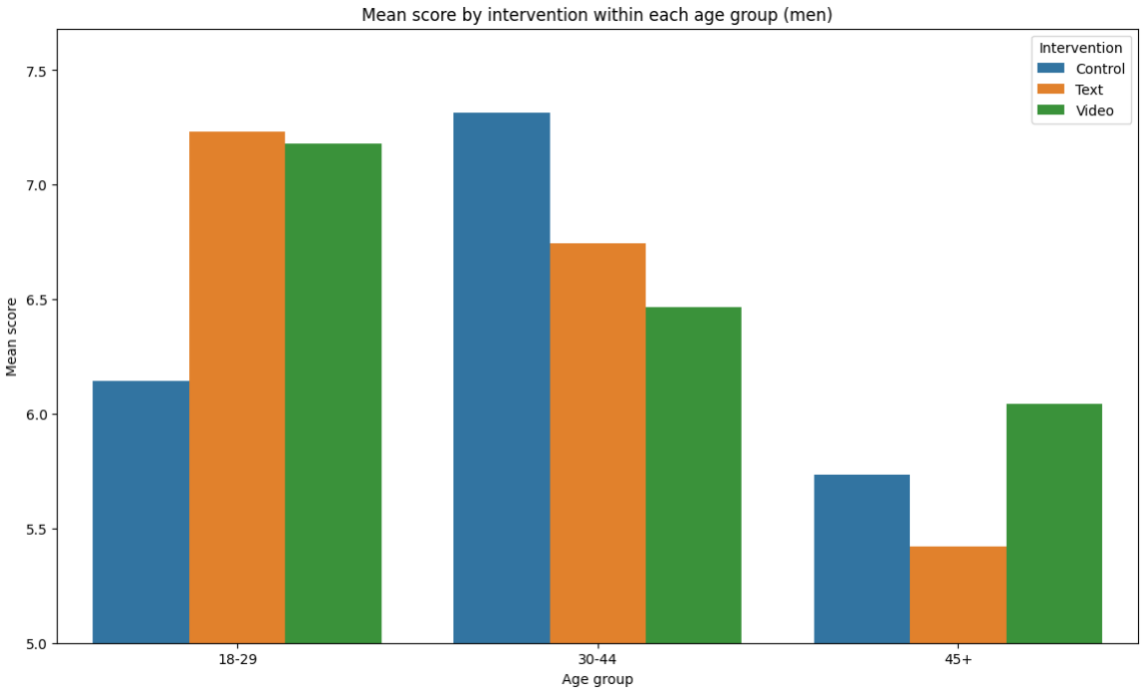

Figure 7: *Mean score by intervention within each age group (men)*

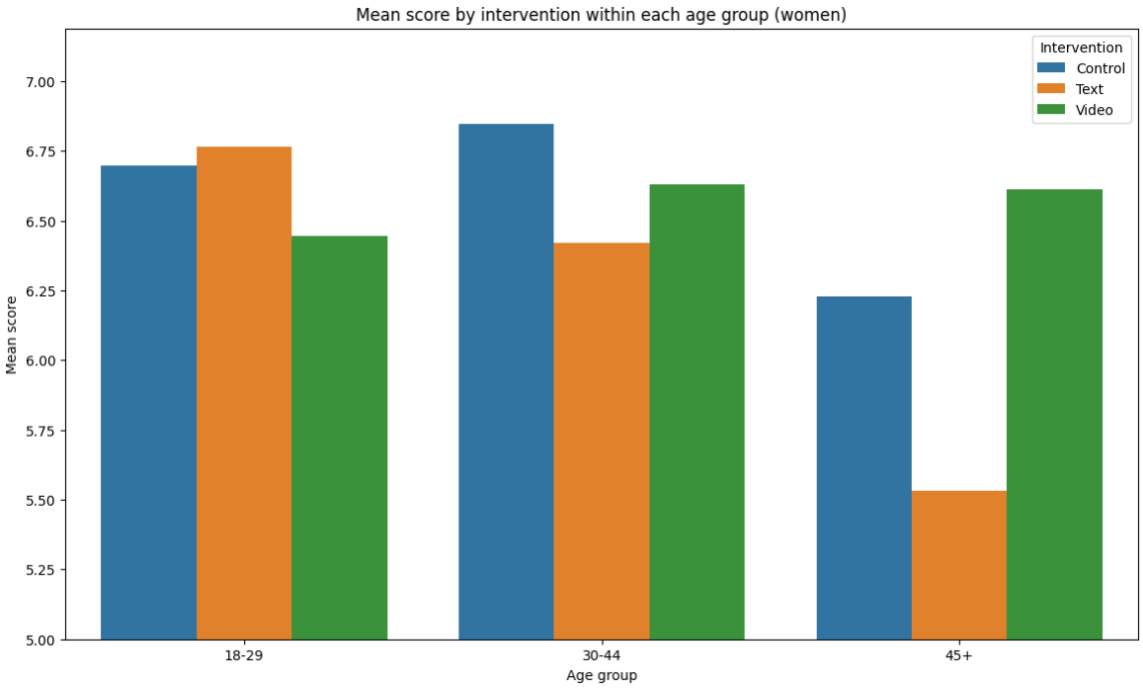

Figure 8: *Mean score by intervention within each age group (women)*

Young men did not score well in the control group, but their mean score increased by an entire point for both interventions. In other words, a 10% improvement. Meanwhile, the middle age group saw a high mean control score on par with the intervention groups for the younger age group. However, interventions caused their scores to drop. The oldest group of men saw trends like women, in that they scored low, text intervention made it worse, and the video intervention made it better. Women saw slightly lower scores for intervention in the middle age group, and little change in the youngest age group. Overall the changes in mean scores for women was not statistically significant ($p = 0.69$) but was for men ($p = 0.03$).

Additional Analysis

Additional analysis examined if other factors were of any influence. One question is whether the amount of time it takes to work through the questions influences the score of the participant.

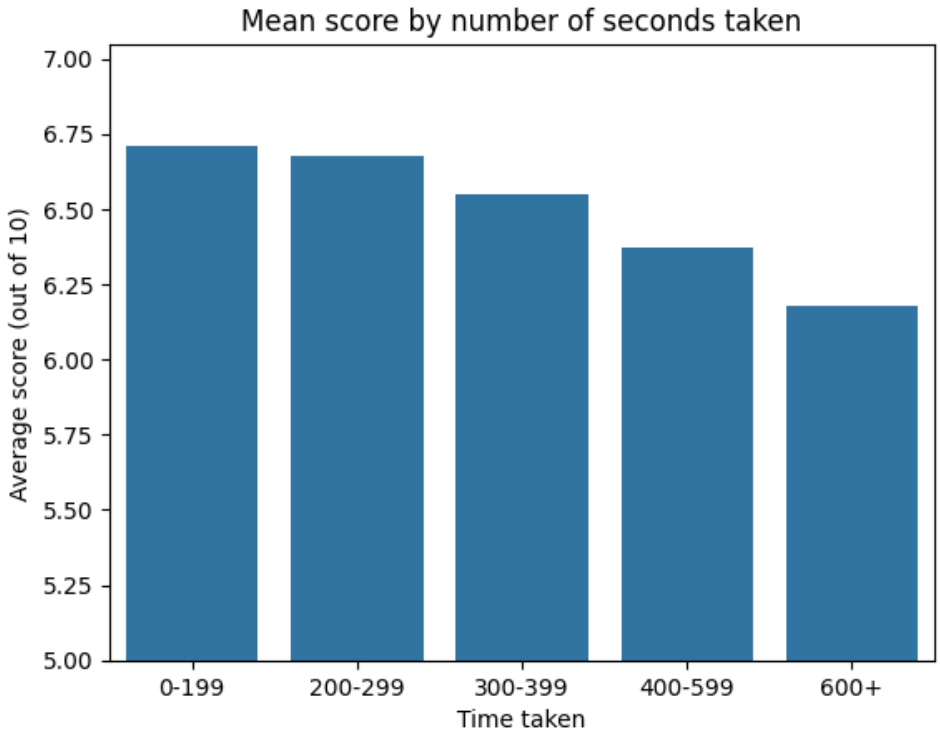

Figure 9: *Mean score by number of seconds taken*

Participant scores slowly decrease the longer they take. The people who took the longest on the quiz did the worst. They also saw the most variation in their scores. This contrasts with the mean accuracy on different questions, which had significant differences in scores between questions.

(no=unaltered, yes=deepfake)

1. No 2. No 3. No 4. No 5. No 6. Yes 7. Yes 8. Yes 9. Yes 10. No

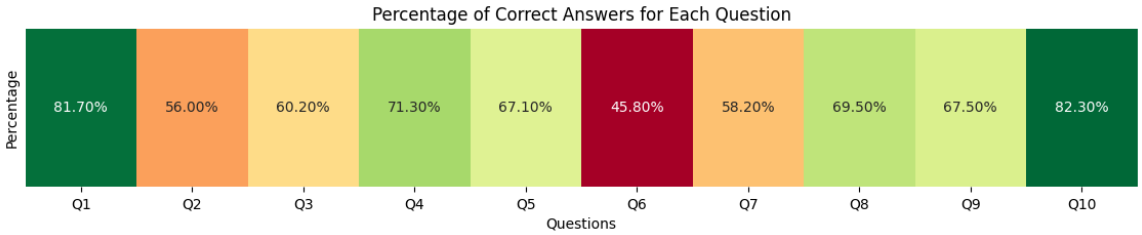

Figure 10: *Percentage of correct answers for each question*

The worst was question 6, where only 46% of participants correctly stated that it was a manipulated video. In contrast, participants did best with the first and last question, correctly guessing 82% of the time that was an unaltered video. This wide range of accuracy demonstrates the wide range of quality when it comes to deepfakes.

V. Discussion

The findings of this study reveal nuanced insights into the efficacy of interventions aimed at improving human recognition of deepfakes across various demographics. The differential responses to interventions between younger and older participants, as well as the slight variations observed between genders, underscore the complexity of designing universally effective educational measures. These results echo and extend upon existing research that has identified demographic factors as significant in technology adoption and media literacy skills (Caramancion, 2021; Nightingale et al., 2020).

Interpretation of Results

Younger Participants' Positive Response to Interventions. The positive response of younger participants to both textual and video interventions might be attributed to their native exposure to digital media, enabling them to discern nuances in digital content more effectively.

Adverse Reaction by Older Participants. The adverse reaction of older participants to interventions, particularly textual ones, and their generally lower performance in recognizing deepfakes, highlights the challenge of media literacy among older populations. This finding corroborates studies indicating that older individuals may struggle with digital literacy and are more susceptible to misinformation (Caramancion, 2021). The observed deterioration in recognition capabilities upon intervention suggests that the interventions might inadvertently increase skepticism or confusion, emphasizing the need for tailored educational approaches.

Connections to Existing Research

The observed effectiveness of interventions among younger individuals but not among older adults suggests a potential gap in existing digital literacy and media education frameworks. Previous research has emphasized the importance of cognitive factors and familiarity with digital technologies in detecting manipulated content (Bronstein et al., 2020; Pennycook & Rand, 2019). These findings extend this research by highlighting the need for interventions that are not only cognitively engaging but also accessible and relevant across different age groups.

Implications for Policy and Education

The implications of these findings are significant for educators, policymakers, and technology developers. For educators, the results underscore the importance of integrating digital literacy education that accounts for demographic differences, particularly age, into curricula. For policymakers, the findings highlight the need for supporting media literacy initiatives that target vulnerable populations, such as older adults, to safeguard against the societal impacts of deepfakes. Technology developers might consider these insights in designing user-friendly tools and resources for deepfake detection that cater to a broad user base.

Limitations

This study was limited in several factors. First was the small sample size of approximately 500 participants that were limited to English speaking US adults. Only two interventions were explored. The study focused exclusively on video deepfakes. The study was also limited to samples from the celeb-df dataset.

Future Research Directions

Expanded survey reach. Future research should explore a wider variety of interventions with a broader audience. Researchers could explore deepfake recognition with children, non-English speaking participants, or international individuals.

Exploring a wider array of deepfake technologies. As deepfake technologies continue to evolve, researchers need to stay abreast of the latest developments and adapt detection methods accordingly. The existing tools for traditional video deepfakes continue to rapidly evolve. Beyond those tools, additional areas of generative AI create new opportunities for spreading misinformation, such as audio, imagery, and text to video generation. Each of these areas can be explored by researchers. Some studies have already found that changing the medium of deepfakes can greatly impact their efficacy (Groh, 2022). Adjusting the available information by modifying or eliminating elements of audiovisual media can change how believable or detectable a deepfake is, highlighting the importance of tailoring detection methods to the specific characteristics of each format.

Public policy development. The increasing prevalence of deepfakes raises important ethical and legal questions. Researchers should consider exploring the implications of deepfake technology for privacy, consent, and freedom of speech. The unauthorized use of an individual's likeness, voice, or personal information in a deepfake can have serious consequences, potentially violating privacy rights and causing reputational harm. Additionally, the development of regulations and policies that can balance the potential benefits and harms of deepfakes will be crucial to ensure responsible use of this technology. By engaging in interdisciplinary research that incorporates perspectives from law, ethics, and technology, researchers can help to inform the development of

comprehensive policies and guidelines that protect individual rights while fostering innovation in digital media. These efforts will be essential to navigate the complex landscape of deepfake technology and its implications for society.

VI. Conclusion

In this thesis, we have explored the growing challenge of deepfakes and whether we can improve an individual's ability to recognize a deepfake. Through a carefully designed IRB-approved survey, we investigated the efficacy of different interventions to enhance human ability to recognize these sophisticated digital fabrications. Our findings reveal a nuanced landscape where younger individuals show a promising ability to discern deepfakes, particularly with targeted education, while older demographics demonstrate vulnerabilities that could be exacerbated by misinformation. Future research should aim to expand on these findings, incorporating a wider array of intervention strategies and exploring the impact of varying deepfake qualities more thoroughly.

The complexity of combating deepfakes demands an interdisciplinary approach. It is clear from our research that no single field can address this issue in isolation. Collaboration across computer science, psychology, media studies, and policy is crucial to developing effective detection tools, educational programs, and legislative frameworks. Such synergy can lead to the creation of robust, accessible technologies that empower individuals to critically assess digital content, regardless of their background or technical expertise.

Moreover, our findings highlight the critical role of education in mitigating the effects of deepfakes. It is imperative that digital literacy becomes a cornerstone of educational curricula, from early schooling through to adult education. Initiatives should not only focus on recognizing deepfakes but also on fostering a critical understanding of media, enhancing the public's ability to navigate the complexities of digital information critically.

This thesis underscores the urgent need for a multifaceted approach to improve human recognition of deepfakes. As we navigate this ever-evolving digital landscape, our adaptability, commitment to education, and collaborative spirit will be our most valuable assets. Together, we can strive towards a future where the integrity of digital media is safeguarded, and the public can engage with information confidently and critically.

References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: A compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2020). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. Journal of applied research in memory and cognition, 9(1), 108-117.

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665.

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. PLOS ONE, 16(6), e0253717. https://doi.org/10.1371/journal.pone.0253717

Caramancion, K. M. (2021). The Demographic Profile Most at Risk of being Disinformed. 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 1–7. https://doi.org/10.1109/iemtronics52119.2021.9422597

Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Aff., 98, 147.

Fallis, D. (2020). The Epistemic Threat of Deepfakes. Philosophy & Technology, 34. https://doi.org/10.1007/s13347-020-00419-2

Garrity, Micheal S., *Voter Suppression Robocall Complaint to Election Law Unit | News Releases | NH Department of Justice*. (2024, January 22). Www.doj.nh.gov. https://www.doj.nh.gov/news/2024/20240122-voter-robocall.html

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1). https://doi.org/10.1073/pnas.2110013119

Groh, M., Sankaranarayanan, A., Lippman, A., & Picard, R. (2022). Human Detection of Political Deepfakes across Transcripts, Audio, and Video. ArXiv:2202.12883 [Cs]. https://arxiv.org/abs/2202.12883v2

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. (2021). Deepfakes Unmasked: The Effects of Information Priming and Bullshit Receptivity on Deepfake Recognition and Sharing Intention. Cyberpsychology, Behavior, and Social Networking, 24(3), 194–202. https://doi.org/10.1089/cyber.2020.0149

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice - People cannot detect deepfakes but think they can. IScience, 24(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Korshunov, P., & Marcel, S. (2020). Deepfake detection: humans vs. machines. ArXiv:2009.03155 [Cs, Eess]. https://arxiv.org/abs/2009.03155

Parkin, S. (2019). The rise of the deepfake and the threat to democracy. The Guardian, Jun 22.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. Arxiv.org. https://arxiv.org/abs/1909.12962

Lyu, S. (2020). *Deepfake Detection: Current Challenges and Next Steps*. IEEE Xplore. https://doi.org/10.1109/ICMEW46912.2020.9105991

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. American Criminal Law Review, 56, 197.

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2020). Can people identify original and manipulated photos of real-world scenes? Cognitive research: principles and implications, 3(1), 1-22.

Pennycook, G., & Rand, D. G. (2019). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings. Management Science, 66(11), 4944-4957.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020).

Grad-CAM: Visual explanations from deep networks via gradient-based localization.

International Journal of Computer Vision, 128(2), 336-359.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2387-2395).

Tidler, Z. R., & Catrambone, R. (2021). Individual Differences in Deepfake Detection: Mindblindness and Political Orientation. Proceedings of the Annual Meeting of the Cognitive Science Society, 43. https://escholarship.org/uc/item/9g67t85v#main

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). https://doi.org/10.1177/2056305120903408

Appendix 1 – Participant Instructions

Intervention 1

You will view 10 videos for 5-15 seconds each. You will then respond if you believe the video was manipulated or not. The distribution of manipulated and un-manipulated videos may or may not be 50/50. Please mark that you are ready to continue when you are done reading this.

Note: there is no audio in the following videos.

Intervention 2

Please read the following information which will provide relevant information relating to the questions.

Some of these videos have been manipulated using 'deepfake' technology. The Oxford dictionary defines a deepfake as:

> "A video of a person in which their face or body has been digitally altered so that they appear to be someone else, typically used maliciously or to spread false information."

Deepfakes includes both altering existing videos (such as changing their lips and voice to say something else or swapping their face to appear as someone else) or using motion capture that is copied to a virtual 'puppet' that pretends to be someone else.

Deepfakes pose a dangerous threat that can be used to spread misinformation and confusion.

Deepfakes can sometimes be spotted through obvious mistakes, such as mismatching skin colors or incorrect muscle and eye movement. They can also be detected through the

'uncanny valley effect', where you can't identify what is wrong but the video 'feels' wrong due to the brain subconsciously detecting errors in the changes.

Please do your best to spot the deepfakes in the following videos.

Intervention 3

This video can be seen at https://www.youtube.com/watch?v=jsLHsEYfPDE.

Appendix 2 – video links

Note: Question order was randomized for participants.

Q 1 - https://youtu.be/fheNpd1CPqk - Real
Q 2 - https://youtu.be/2thOYv1x7mg - Real
Q 3 - https://youtu.be/7WMXtbCKtdc - Real
Q 4 - https://youtu.be/e90EpDFe16I - Real
Q 5 - https://youtu.be/sKb-7d_7ykw - Real
Q 6 - https://youtu.be/hSIffdSzeoA - Deepfake
Q 7 - https://www.youtube.com/shorts/LrowPJPyJ4s?feature=share - Deepfake
Q 8 - https://youtu.be/o1hsFisFSRg - Deepfake
Q 9 - https://youtu.be/iYdFUAfgwUg - Deepfake
Q 10 - https://youtu.be/RzWNSQSXrlg - Deepfake