Discipline of Digital Arts and Humanities
University College Cork

Digital Humanities and Information Technology

(DH4003): (Researc)

Title of the Assignment: Peoples Perception of Deepfakes

Final Year Project 2025

Student Name & Number: Cathal Mccarthy / 121380241

Date: 17/04/2025

Lecturer: (David Murphy)

In submitting this assignment I confirm that the submitted work is entirely my own original work, except where clearly attributed otherwise, and that it has not been submitted partly or wholly for any other educational award.

I hereby declare that:

• this is all my own work, unless clearly indicated otherwise, with full and proper accreditation;

• with respect to another's work: all text, diagrams, code, or ideas, whether verbatim, paraphrased or otherwise modified or adapted, have been duly attributed to the source in a scholarly manner, whether from books, papers, lecture notes or any other student's work, whether published or unpublished, electronically or in print.

Table Of Contents

Abstract

This project explores how well people can detect deepfake videos without any previous training or support. The idea was inspired by my work placement last year, where I was working in the information security sector and was asked to create a deepfake. At the time, I couldn't do it but the request stuck with me and got me thinking about how convincing deepfakes have become and how easily they could be misused. For this study I created a short survey using a mix of real and deepfake videos and asked participants to decide which were real and which were fake. Most people were able to recognise the deepfakes, but there were also a few real videos that people incorrectly flagged as fake, showing a level of doubt and uncertainty. The project highlights how tricky deepfake detection can be even when people are aware of the risks. It suggests that while instincts are sometimes accurate, people could benefit from better tools, training and awareness. The findings contribute to the wider discussion around media trust, misinformation, and the role of human judgment in a world where seeing is no longer always believing.

Introduction

In today's digital world, it's getting harder and harder to tell the difference between what is real and what is fake especially when it comes to videos. Deepfakes being the main reasoning behind this statement. Deepfakes are videos which are made by individuals using Artificial Intelligence (AI). These are videos where someone's voice, face or movement is altered or where a voice or a action is swapped realistically to make it look like something they said or did something they did not actually do. There has been a massive change in deepfakes over the last number of years making them way more common as they are way more accessible. This is evident from this research project as I was able to make these deepfakes through an app which I found online meaning that most if not everybody has access to create these deepfakes. Like most things in technology with the positives for films and all that kind of industry there also a lot of negatives around the trust people have in videos now, misinformation which is a big one and digital safety.

This project "People's Perception of Deepfakes" explores how well people can spot a deepfake. The main questions I was focusing in on and what I wanted to find the answers to were the following, can we still tell when something is fake just by looking at it? And if so what gives it away? The aim for me was I wanted to see what stood out for people watching these videos what exactly caught their eye and what stood out to them and how confident they were in their final decisions. My short online survey which gave people real and deepfake videos gathered information on peoples understanding on deepfakes. It gave me a good detection how real deepfakes could and couldn't be and along with that what were the cues that made a video suspicious for example. It's a mix of psychology, media studies, and digital tech so it fits well under the umbrella of Digital Humanities and Information Technology.

Why This Project Matters

Deepfakes have transformed from being a tech novelty to being something we need to look after and find a solution for. Deepfakes are being made in every of aspect of life from fake celebrity videos to altered political speeches, there really are no limits when it comes to deepfakes. The fact that anyone with access to a computer and any bit of ambition, can easily make a convincing fake video has a lot of big implications for online safety and people's trust.

We've already seen deepfakes used in scams and misinformation campaigns, and as the technology improves, it's likely we'll see even more of this.

What makes this even more complicated is how realistic deepfakes are getting. Some are after getting so realistic that even experienced viewers along with detection systems struggle to determine whether the video is real or fake. This brings up some big questions that cannot be answered now.

If AI can create something so believable, how do we protect ourselves from being fooled?

Can humans still trust what they see online?

How do we decide what's true?

While a lot of work is being done on training different detection systems to detect deepfakes there has been very little focus and work done on how well people can detect deepfakes. That's where I got the idea to make this project, I wanted to test different videos with regular viewers and see how they judged or reacted to simple deepfake videos that I made myself. The main thing was they had no training they just used their own instincts.

How the Project Was Carried Out

As a test to see how well people can detect deepfakes, I designed and launched a short online survey that presented participants with both real and AI-generated videos. The app I used to create the Deepfakes was called Synthesia. This app gave me the freedom to create the videos I wanted to create using ai avatars that are surprisingly realistic. The app itself was very easy to use. That reflects just how available this technology is to the general public.

The survey I built was on Qualtrics, which let me collect responses anonymously and in a structured format. Participants were asked to view 6 short video clips and decide if they were real or fake. After each clip, they were asked to explain their thinking and rate their confidence. This process gave me useful data on how people react to deepfakes and what stands out to them.

The project had three main stages:

1. Creating the videos and testing them
2. Designing and distributing the survey
3. Analysing the results to look for patterns in decision-making and confidence levels

Why This Project Fits

This project is a perfect example of what the DHIT course has been all about over my last 4 years of studying it. It brings together digital tools, media literacy, and the human side of technology. I was able to use creative tools to build something new, while also researching its impact and collecting real human feedback.

Creating the videos, building the survey, and analysing the data gave me a chance to apply everything I've learned over the past few years from understanding digital culture and ethics to using tech tools in a meaningful way. It's not just a study on deepfakes it's about how humans interact with emerging technology and how we make sense of what we see.

Literature Review

Deepfake technology has developed rapidly, generating concern on serious ethical concerns and security risks in various domains. For instance, media and politics, and cybersecurity. It is produced by the drive of artificial intelligence. Generative adversarial networks (GANs) produce realistic manipulated videos. Its application in the creation of deceptive media puts public trust, the control of misinformation, and even national security at risk. Among the fundamental issues is use in disinformation operations. Deepfakes in politics produce fake speeches or political leaders behaviours. It can impact elections (Chesney & Citron, D. (2019). It undermines democratic processes and further erodes public trust in journalism. The technology is also applicable in cybercrime, identity theft, and financial misrepresentation.

The photorealistic synthetic media can deceive people into believing false reports. This heightened risk has spurred enhanced effort on the part of governments, tech companies, and scientists to develop effective detection tools. The human ability to tell authentic from manipulated material is still a cause for concern. Human perception is relative to various variables. Knowledge of the subject and subtle cues buried in deepfakes are only a couple of

examples. Studies show unnatural motion and more unreal expressions as a few crucial signals that might allow humans to detect synthesized material (Chesney & Citron, 2019). But with technological advancement, these indicators become harder to identify. So, relying solely on human instincts is insufficient for detecting forgeries.

The objective of this study is to test the efficacy of humans in deepfake detection. The emphasis of the study on perceptual information used by participants in discriminating deepfake and real videos. It involves creating a controlled data set of progressively manipulated deepfake videos. Then analysing participant ratings for their detection efficacy. The study seeks to contribute towards the overall understanding of human perception of deepfakes. This will provide insights into optimizing detection mechanisms.

The Evolution of Deepfake Technology

Deepfake technology has been driven by artificial intelligence. This is particularly in computer vision and deep learning. Generative adversarial networks (GANs) are an artificial learning model that was invented by Ian Goodfellow in 2014 (Vaccari & Chadwick, 2020). GANs places two neural networks in a battle with each other—the generator. This creates artificial content, and the discriminator, which attempts to find out if it is real or not. Through data training, the model produced more realistic images and videos with every iteration. This formed the foundation for today's deepfake technology (Ajder, 2019). When deepfakes first emerged, the technology was less sophisticated. It results in noticeable artifacts like unnatural facial expressions, stretched facial features, and unpredictable lighting. However, as machine learning algorithms improved, generation methods developed a great deal more.

Alanazi (2023) identified the great level of progress into AI-based face manipulation. To mention a few, there include facial re-enactment and lip synchronization advancement, expression transferring. These kinds of technologies contributed to the real production of a highly realistic video deepfake hard to distinguish it from the natural video.

Massive quantities of training data also contribute to the success of deepfake technology. Training data is split into much smaller portions, which are utilized to allow machine learning algorithms to learn and get better at creating realistic content (Vaccari & Chadwick, 2020). DeepFaceLab and Face Swap are open-source deepfakes. They provide developers and researchers with the tools that they need in order to experiment with and develop new deepfake techniques. Furthermore, social media and video-sharing websites have also become sources of publicly available facial information.

They can be used to train deepfake models more effectively (Wasteland, 2019). It has become increasingly hard to tell apart authentic and fabricated media due to the development of deepfake technology. Detection previously involved the detection of visual artifacts and inconsistencies such as unnatural blinking or facial feature misalignment. But, with more advanced algorithms, these restraints have been bridged to a large extent. This makes outdated detection mechanisms irrelevant. Scholars have created AI-driven detection models that analyse millisecond facial expressions (Vaccari & Chadwick, 2020). For example, micro-expressions and physiological responses, in order to detect the authenticity of a video.

Advances in deepfake generation processes are an ongoing cause for concern for detection and countermeasures. The technology advancement has introduced unprecedented ethical and security concerns. Their application in disinformation campaigns and privacy intrusions have triggered international debate about regulation. Governments and technology firms are funding research so that they can combat hazardous deepfakes. But the sheer speed of innovation means it is difficult to stay ahead of new threats.

Ethical and Security Concerns

Deepfake technology has also been the subject of intense debate concerning its ethical implications. The potential of the technology to create very realistic but false media has sparked concerns. This has been especially in political, social, and security contexts (Alanazi & Asif, 2023). The technology, over the years, has become more sophisticated. Synthetic

media is now virtually impossible to tell apart from real footage. The advancement has turned the technology into an effective tool of deception with cataclysmic consequences. Chesney (2019) argues in the context of broader disinformation warfare about deepfakes. It demonstrates the potential that it can undermine democratic processes.

Therefore, the capability of it to deceive large groups is gravely jeopardizing the free operation of democratic institutions. The technology can generate deceptive speeches, interviews, and live performances of political figures (Alanazi & Asif, 2023).
This has the potential to manipulate public opinion. In this case, the public will be deceived into believing false narratives in favour of some political agendas. An example is that the technology can be used to discredit candidates or create scandals. It impacts the attitudes and behaviours of voters. Its potential risk to democratic transparency and integrity renders them a top priority in today's fight against digital disinformation. Vaccari (2020) also discusses the dissemination of false information through deepfakes, particularly on political levels. Some of the real situations where manipulated media contributed significantly to public opinion and social interactions are numerous.

One of the means through which the technology has been abused is in creating fake videos of politicians. This has either been achieved by misrepresenting their activities or placing them in a misleading context. The videos have, in some instances, gone viral. They have millions of views before they can be established as false. The rate at which deepfake videos spread on social media makes it difficult to prevent them from spreading. This is because disinformation spreads widely before fact-checkers or authorities can act against it. Additionally, the emotional nature of visual media—particularly video—makes manipulated media more potent that sway public opinion. Deepfakes also create ethical concerns outside politics. Deepfakes have also been used for other cybercrimes (Westerlund, 2019). For example, fraud, identity theft, and blackmailing.

The technology also has implications deeper in the world of cybersecurity and law enforcement. The ability to generate photorealistic, synthetic video makes deepfakes a valuable tool for fraud and cybercrime (Westerlund, M. (2019). It can be taken advantage of,

for instance, by bullies to impersonate a company manager on an internet conference and convince employees to send large sums of money. Such incidents would have severe fiscal as well as reputational impacts. Similarly, the technology can be used to fake evidence in court cases. It can generate false alibis or simulate criminal behaviour. This taints the credibility of court cases. Courts and police are unable to tell between authentic and created media.

Kaate et al. (2023) carried out a study with 46 participants using a think-aloud approach to see how people respond to deepfake personas compared to real human ones. The goal was to understand how users perceive deepfake-generated personas and what that means for Human-Computer Interaction (HCI). The study found five key themes:

Realism

Participants judged how lifelike deepfake personas looked, focusing on facial expressions, lip-sync accuracy, and emotional delivery. While some found them believable, odd movements and expressions often made them feel unnatural, triggering the uncanny valley effect.

User Needs Representation

The study looked at whether deepfake personas could accurately communicate user needs. While deepfakes were more engaging than static text-based personas, some participants felt they lacked the genuine human touch needed to fully represent emotions and concerns.

Distractions

Many participants pointed out distracting elements in deepfake personas, like strange eye movements, stiff facial expressions, and occasional glitches. These quirks pulled focus away from the message and made interactions feel less smooth.

Added Value

Deepfakes have potential in UX design, marketing, and digital assistants, but there are major concerns around ethics, privacy, and the possibility of deception. While they could improve engagement, their use needs to be handled responsibly.

Trust and Connection

Trust played a big role in how people perceived deepfake personas. Some users were open to them, but others were wary because of how AI makes decisions and whether the personas felt truly authentic. The lack of clear disclosure about deepfakes being AI-driven made some people hesitant to trust them.

The study emphasizes the need to improve deepfake realism, fix glitches, and be more transparent about how they're used in HCI. Developers should work on making facial expressions more natural, reducing noticeable flaws, and allowing users to give feedback on their experiences. These changes will be crucial in making deepfake personas more widely accepted.

This research shifts the conversation on deepfakes beyond just misinformation, highlighting their role in digital interactions and UX. It also points out the importance of AI ethics, deepfake detection, and building safeguards to prevent misuse. Future research should explore better ways to measure user perception and study how people engage with deepfake personas over time in real-world settings.

Human Detection of Deepfakes

The ability of human beings to identify deepfakes is a hallmark challenge in the fight against digital forgery. Since the technology has continued to evolve, it has been difficult for the general public to differentiate between real and fabricated media (Chesney & Citron, 2019). While research shows that there are intrinsic visual and audio cues that humans can use to identify manipulated media, these cues are progressively becoming undetectable as

algorithms used to generate deepfakes also improve. Researchers identified a number of characteristics human viewers will use to identify deepfakes. Mechanical smiles, or unnatural facial movement, are typically the initial indication of artificial content (Ajder, 2019). The underlying technology of face manipulation software still has a problem with creating realistic transitions between facial expressions, which leaves noticeable defects. For example, when it creates a smile, the process may not synchronize the eyes and facial musculature overall with the motion of the mouth. To the human eye it may appear unnatural. Similarly, the inability of its algorithms to correctly duplicate complex micro-expressions lead to characteristic tell-tale clues that something is wrong. Asynchronous lip movement is another crucial detection cue.

In the majority of deepfake videos, synchrony between a subject's mouth movement and the sound they are making is less than perfect (Chesney & Citron, 2019). The difference varies from barely noticeable inconsistencies to more obvious errors. The mouth becomes desynchronized with the words that lead to inconsistency. The technologies also fail to exactly mimic the natural movement of the eyes. Studies have also suggested that atypical blinking, atypical eye movement, or evading eye contact can also signify fake content (Westerlund, 2019). In the natural world, the eye movement and blinking behaviour possess some rhythms challenging for AI algorithms to simulate in a natural way.

However, with advancing deepfake technology, it is getting harder to spot these inconsistencies. The use of advanced generative adversarial networks (GANs) has made deepfakes considerably more realistic (Chesney & Citron, 2019). It reduces detectable artifacts such as rigid facial expressions or asynchronous lip-syncing. More sophistication also makes it challenging for casual observers to rely on instincts alone. Therefore, there is a growing need for a more sophisticated way of detecting them. Trained individuals have been shown to be able to identify deepfakes as opposed to untrained viewers. Forensic experts, for instance, have trained to observe minor differences in digital content. For example, poor lighting, pixelation, or irregular shadows. Experts identify deepfakes that an untrained individual would believe. This serves to underscore the fact that training and experience lie at the centre of identifying synthesized media (Alanazi, 2023). Training

modules that allow one to spot deepfakes are becoming more well-known. More detection may be possible among the general population.

However, the more sophisticated nature of newer technology continues to pose a challenge (Chesney & Citron, 2019). Therefore, continued investment in training modules and in media literacy training for the public continues to be needed. There has also been investigation into whether deepfakes can be detected using crowdsourcing. In some research, while single viewers are less capable of detecting them, larger groups of people are more capable of detecting them as a crowd. Crowdsourcing sites, where several viewers watch the same video and then review it, can possibly utilize the wisdom of the crowd.

Challenges in Deepfake Creation and Detection

Both the production and detection of deepfakes are technically and ethically challenging. Both impact studies as well as practical applications of the technology. Perhaps one of the largest obstacles is the ethical and legal limitation on creating deepfakes (Vaccari & Chadwick, 2020). That particularly through use of public figures' faces and voices. Various sites and authorities have implemented protection against exploitation of public figures in misinformation. These include damage, misinformation, and privacy invasion threats. Such limitations pose challenges in creating deepfakes for academic use. It limits the extent of varied and natural data for evaluating detection algorithms. For example, copyright law usage, for instance, for using a politician's or a celebrity's photo, creates legal barriers for academics seeking to produce deepfakes without violating laws. This restriction may hinder research advancement that requires quality deepfakes to attempt detecting algorithms and human responses.

Ajder (2019) notes that even though the technology has come far in its mimicry, there still are issues of consent and copyright law. Applications for evil ends such as defamation or dissemination of false information have generated increasing attention from policymakers as well as social media. Therefore, most researchers are encumbered by the process of obtaining permission to use realistic deepfake data (Vaccari & Chadwick, 2020). This affects

the generalizability of findings. For instance, a study whose aim is to find in videos that include public persons must deal with such legal issues. Therefore, this limited the scope and applicability of findings. Apart from legal and ethical issues, it's a constant war of technology to catch deepfakes.

Even though detection tools have been developed utilizing AI, they aren't perfect and are continually being outdone by evolving technology (Vaccari & Chadwick, 2020). Generative adversarial networks, or GANs, have made deepfakes even more complicated. Infinitesimal artifacts are harder for human vision and AI to spot. Westerlund (2019) describes that the adversarial arms race of creating and detecting deepfakes has resulted in the evolution of adversarial AI methods. In this case, creators deliberately bias algorithms in a way that they cannot be detected. For instance, creators can employ noise addition or pixel pattern manipulation to mislead detection algorithms.

In this case, AI finds it difficult to distinguish synthetic media from real media. Moreover, detection software itself is also facing the problem of scalability. Even though some systems based on AI have fared really well when there were tightly controlled lab environments, they tend to come crashing down once real data is processed (Vaccari & Chadwick, 2020). These systems need to address variations in light, video, and context and how this affect detection accuracy. A controlled-produced deepfake may be easier to detect than a non-controlled-produced one. The environment and the placement of the lighting and cameras would be highly mixed. Detection methods will thus need to get trained on increasingly large numbers of datasets in order to keep up.

Alanazi (2023) argues that detection software powered by AI has to stay abreast with evolving tactics of deepfake creators. Machine learning models have to be trained on multivariant datasets, which specify how deepfakes have to be edited. Despite intensive training, no detection framework can ever be totally fool proof. With the development of technology, it is important that researchers develop detection systems that are resilient and robust (Vaccari & Chadwick, 2020). The systems should be able to detect new forms of manipulation without being rendered outdated by technological development

Contribution of This Study

In endeavours to make a contribution towards the understanding of human sensitivity to deepfakes, the current knowledge will be discussed. As technology keeps improving, the bar has been raised for human capability to discriminate between deepfakes (Chesney & Citron, 2019). Ongoing efforts have been made in developing AI-centric detection systems. However, human perceptual cues are still applicable when it comes to detecting deepfakes. The novelty of this research is in performing controlled experiments that are designed for human detection accuracy. In addition, detection of perceptual cues can be utilized to separate synthetic media from authentic content (Chesney & Citron, 2019). Therefore, this research identifies the specific cues that lead to successful detection. And as such, will gain important perspectives in human communication with digital media.

Recent literature refers to soft visual and auditory cues as being paramount in recognizing deepfakes. Abnormal facial movements, lip-sync issues, and involuntary eye movements (Ajder, 2019; Westerlund, 2019), for example. However, as the technology becomes more developed, the cues are even softer. This is less apparent to non-technical people. For example, research has proven that although facial inconsistencies are a general cue, more recent approaches have reduced the recognisability of these irregularities (Chesney & Citron, 2019). It is a proof that there should be more research on how these cues evolve and affect detection precision. The experiments carried out under this study provide an even more formalized method of establishing the effectiveness of such cues as well as the effect they have on human detection ability.

One of the most compelling arguments of this research is that it deals with detecting deepfakes through human perception and not through AI-based approaches. Even as AI systems improve in their ability to detect deepfakes, there is no guarantee of their success. Chesney (2019) explains the vulnerability of AI adversarial deepfake detection. Here, the creators do the extra step to mislead detection systems. This research seeks to complement AI-based systems by highlighting the role of human detection. It provided insight into how human beings can be trained to identify deepfakes and aid in developing a more robust

detection system. This study is looking to help in the development of better training tools for humans and machines.

The findings of the current study also have considerable ramifications for public campaigns and media literacy. It is critical that people learn how to apply critical thinking skills such that they can judge the validity of electronic information. Vaccari (2020) reveals how deepfakes influence the emotions of individuals and how manipulated content distorts truth and structures social relationships. The purpose of the study is to make individuals more discerning consumers of online content. The results constructed could be utilized in curricula to educate individuals regarding the danger of deepfakes.

Conclusion

The technology of the deepfakes has evolved a lot in recent years. This is especially in putting media integrity, security, and ethics at risk. The more sophisticated the technology, the more challenging it is to distinguish between actual content and generated content. This poses immense threats to social stability and public trust. As good as the detection tools have become, these technologies are not sufficient to keep pace with the ever-evolving nature of deepfakes. The limitation of AI-driven solutions calls for the need to bring human perception into detection. This work will attempt to advance our understanding of how humans perceive and detect deepfakes. It will mainly focus on the subtler cues' humans use to estimate genuine vs. synthetic videos. This project will identify prevailing perceptual patterns and accuracy levels. Therefore, it will give substantial insights into what determines human detection capability. The outcome will also be fed back into more effective training procedures. Therefore, individuals will have the capacity to detect deepfakes and withstand disinformation. Moreover, the study will inform the creation of countermeasures against internet fraud, to minimize the negative impact of deepfake technology on public discourse. The study hopes to empower individuals to analyse the authenticity of digital information.

Analysis

What's the Problem?

The problem I'm looking into with this project is how good people actually are at spotting deepfake videos without any kind of training or tools. A lot of work is being done on developing AI systems to detect deepfakes, and some of them are very advanced. But there hasn't been nearly as much research into how everyday people react when they see a deepfake and whether they can actually spot one or not (Vaccari & Chadwick, 2020).

The technology has come a long way which will start to cause an issue sooner rather than later. Deepfakes used to be easy enough to catch out with weird facial movements, bad lip-syncing, or lighting issues. But now with advancements in technology and a lot more time they look way more realistic which makes it harder for people to tell the difference. These videos are now around people whether they like it or not social media apps messaging apps and even sometimes in the new. This allows very little time for a human to make up their mind based on what they have just seen. This leaves a massive problem or hole in society if people can't recognise what is real or fake anymore leaving serious trust issues. It affects trust, media literacy, and even how people engage with the world around them (Chesney & Citron, 2019; Westerlund, 2019).

What Am I Trying to Find Out?

The main objective I am looking for in this research project

Can people with no training or background in AI actually spot a deepfake when they see one?

I also want to understand what gives it away to people

What kinds of clues give it away to the general public that these videos are fake ?

Do people look at the eyes, facial expressions, how natural someone's voice sounds or are they just guessing based on gut feeling?

From reading around the topic my hypothesis is:
Most people won't be able to consistently tell which videos are deepfakes, and when they do make a choice, it'll be based on things like awkward facial expressions, eye movement, or bad lip-syncing. One thing that might alter my hypothesis is the fact that they will know beforehand that there are deepfake videos in the survey and also the fact they will be compared to real life videos I have taken off YouTube.

Research shows that people often rely on these kinds of visual clues, especially if something feels off or it just doesn't look right. (Ajder, 2019; Alanazi & Asif, 2023). As of a result of the improvements to deepfakes and deepfake apps I believe the small little things like lip syncing etc will be way harder to notice. That means the cues people use might not even be that reliable anymore which makes detection even tougher (Westerlund, 2019).

Why This Project Is Important

Even though AI tools are getting better day by day at spotting and recognising a deepfake there is still a high reliance in people being able to determine if a video is real or not. It is still the people who share the videos across social media or through word of mouth and ultimately it is humans who make up their mind what to believe and what not to believe. If we don't understand how people actually judge this kind of content it is hard enough to know where the risks lie with deepfakes. This project was designed not entirely to build a detection system but really to understand how people make their decisions in the first place. That is why I think my project is useful and beneficial.

This study looks at how people react to deepfakes in a casual setting just watching short clips and giving their opinion based purely on the video they have just seen. It is very similar to how people would come across a video online apart from the fact that they know 3 out of the 6 are fake videos. The insights into their answers could help me in the future with training different awareness campaigns or even educational tools that might help viewers in the

future. These will all stem from the responses I get from my study how confident people are with their answers and what gave it away etc.

Design

Overview of Experiment Design

The design of this study was shaped around the goal of analysing how accurately untrained viewers could detect deepfake videos in a controlled setting. To do this I created a short digital survey that presented the participants with a series of 6 videos 3 being real videos the other 3 being deepfakes which are all Ai generated. The idea was to simulate the kind of instinct-based decision making that happens on all types of social media when people are scrolling through videos online with a little added structured approach. Each video was accompanied by 5 different questions which allowed me to gather meaningful data with such basic questions. Participants were asked to identify whether they believed the clip was real or fake, explain what influenced their decision, and rate their confidence level. (Kietzmann et al., 2020).

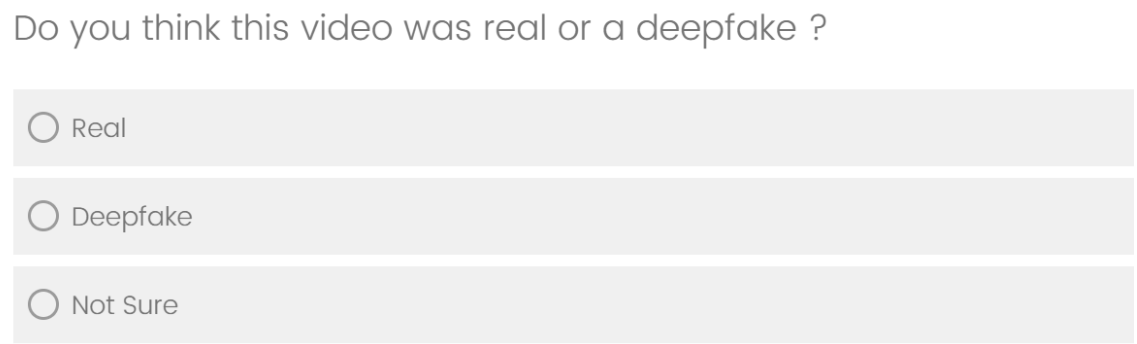
Video Creation and Preparation

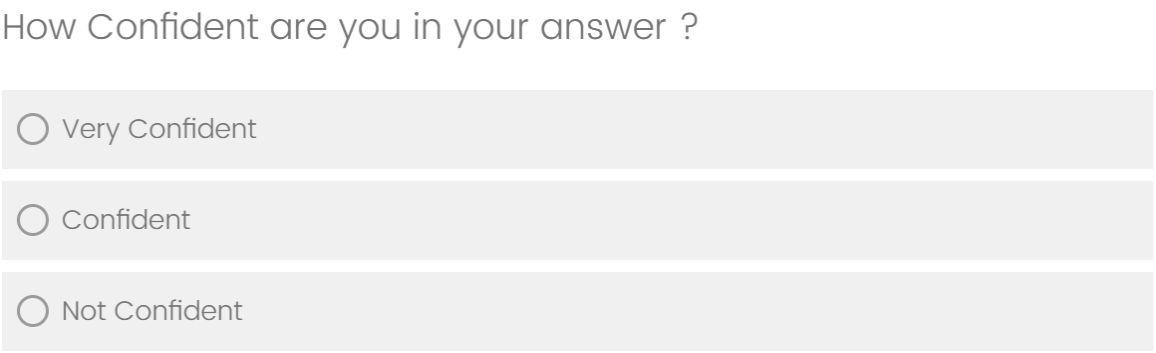
To carry out the before mentioned experiment, a set of six short video clips were created 3 being real and 3 being deepfakes. The deepfakes were generated using Synthesia, a popular AI-based video creation tool that allows users to produce realistic Ai avatar led videos with voice and facial syncing. The platform was chosen for its accessibility and ability to produce high quality content quickly which shows how easy and available these tools are to the general public. Many other different Ai based creation tools were trailed in the process but failed to carry out small different aspects that Synthesia provided. After generating the AI videos, additional editing was required as I found the start and end of the generated images looked too fake there was a second either side of the talking / content which gave it away completely. I used Adobe Express Video Trimmer the same video trimmer I used for the YouTube clips in order to complete this. Small little features like added captions and

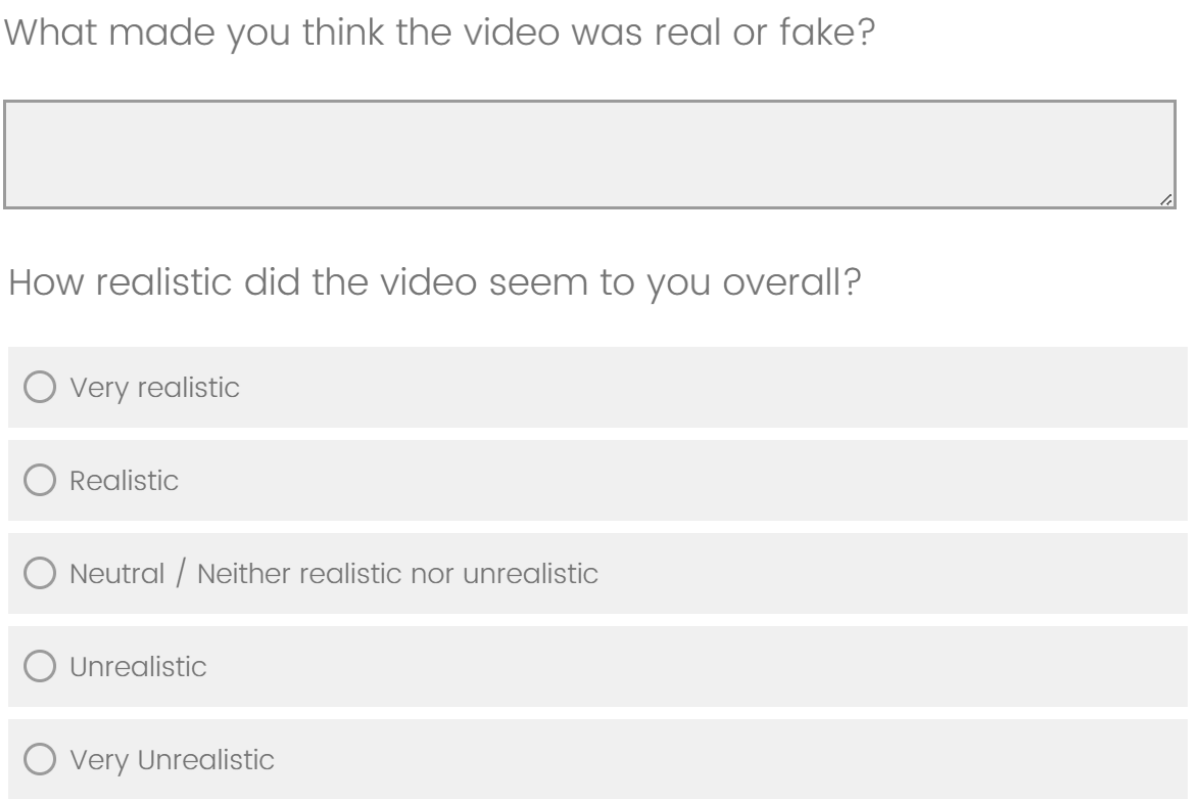
background noise were also implanted into my Deepfake videos in order to make them look more realistic. These changes were made to ensure that differences in quality or presentation didn't influence participant judgment, and that any guesses were based on the content itself, not obvious visual clues or interface elements.

Survey Structure and Question Design

The videos were embedded into an online survey created using Qualtrics which was chosen for its flexibility in formatting and ability to collect both quantitative and qualitative data. Each video clip was followed by five key questions: a yes/no judgment on whether the video was real or fake, an open-ended question asking participants to explain what made them feel that way, a confidence rating on the video , a how realistic question asking the user how realistic the video seemed to them and finally a question asking the user what influenced there decisions the most with a list of the most common standouts of deepfakes from the papers I have read. Below are the questions I used in my Survey.







This combination was designed to not only test accuracy but also capture the reasoning behind each decision and how sure the participants felt about their answers. By collecting both numerical responses and written explanations, the survey aimed to explore both the instinctive and reflective sides of human deepfake detection. This structure is supported by previous research, which emphasises the value of combining binary classification tasks with

open reflection to better understand how people engage with synthetic media (Groh et al., 2021).

Tool Selection and Challenges

One of the more unexpected challenges in the design phase was finding a suitable app to create the deepfake videos. I tested several tools, including both mobile apps and browser based platforms, but many of them didn't meet the requirements I had in mind. Some lacked customisation options, others didn't allow me to upload my own scripts, and a few had limited facial realism or unnatural movement. A few tools were also very general aimed at entertainment rather than realism which made them less useful for a study focused on detection. After quite a bit of trial and error, I chose Synthesia because it allowed me to input custom scripts, choose from a variety of realistic avatars, and create consistent-looking videos that suited the tone of the survey. This process helped reinforce how accessible deepfake tools have become, but also how difficult it is to find one that strikes the right balance between quality, control and realism for research purposes.

Participant Selection and Ethics

Participants for the survey were gathered informally through people I knew mainly friends, classmates, and family. This helped me get a bit of variety in terms of age and background, even though the goal wasn't to have a massive or perfectly balanced sample. Instead, my goal was to see how people would react to the same videos without any specific deepfake training. I only asked for basic info like age range and gender to keep things simple and make the survey quick to fill out. Everyone was told that their answers would stay anonymous and were just being used for my final year project. Ethics approval was granted by University College Cork before the survey went live. At the start of the survey, I added a consent forma that all participants were shown and in order to complete the survey they had to give there consent. I followed standard ethics guidelines for low-risk online studies, where keeping things private and making sure people agree to take part properly is really important.

Summary

Overall, the design of my test was aimed to be very simple and also natural for the user while also collecting useful data that will help me analysis the data properly. The thought process behind the short, quick and realistic videos in the survey was to replicate the videos the average normal person would be seeing on the outside world through different social media apps etc. The idea with the open text response was that it gave users a voice to give their own opinions and reasons for considering it real or a deepfake. The other options then multiple choice was there to get the best possible data from the conducted survey. As of a result of the survey being built around the person doing instinctive reaction rather than a through analysis it suited the aim of my project perfectly to test how regular people judge deepfakes organically without any training etc. This design allowed for both individual differences and shared patterns to be captured, which could be useful for future research or public awareness campaigns.

Implementation

The survey was created and hosted on Qualtrics meaning it was very easy to embed the videos into the survey, collect responses and it made it very easy to collect the responses and there corresponding data afterwards. Once the final version was ready, I shared the survey link through an anonymous link which Qualtrics provided through email. This allowed me to reach a mix of people quickly, without needing a formal recruitment process. The survey was left open for just under a week giving people a sufficient time to complete the short 5-minute survey without any time pressure. All the videos were embedded directly in the survey so that participants could watch each clip and click the follow-on button which then allowed them to answer the questions directly afterwards. This helped keep the experience smooth and straightforward like how someone would naturally consume short-form video content online.

Data Collection

As responses came in the app itself Qualtrics automatically stored and updated all the relevant responses to their corresponding questions. This made it easy to track how many

participants had completed the full survey and to check that all the video responses were recorded properly. Once the survey had been closed successfully, I exported the data to Excel so I could separate the different types of answers the deepfake/Real responses, the confidence scores, the authenticity of the videos, the open-ended written explanations and the specific signs that influenced your decision. In some cases, I had to clean up the text responses slightly such as fixing typos or removing vague comments that didn't relate to the videos at all. There were also 1 or 2 attempts where my answers from previewing the survey prior to releasing it were in the results. For the open answers, I also grouped similar responses together to identify common patterns for example if several people mentioned "blinking" or "mouth movement" as a reason for thinking something was fake. This helped prepare the data for the deeper analysis stage later.

Challenges and Adjustments

Overall, the survey ran smoothly enough but there were a few small issues I had to deal with along the way. One of the main ones was making sure the videos worked properly across different devices. Some participants were using phones or tablets, so I tested the survey in advance to make sure the video playback and captions worked correctly on all screen sizes. In a couple of practises runs I did the videos failed to load but when I reloaded the Qualtrics app they ran smoothly again. I had to add the above message to everybody I sent the survey link to. To ensure that all the questions were answered I added a requirement meaning that each question had to be answered before moving on. These small adjustments helped improve the quality of the responses and made sure the data collected would be useful for later analysis.

Summary

By the end of the data collection period, I had gathered a full set of responses that included both clear deepfake or not decisions and a wide range of written feedback from participants.

The videos had been successfully presented in a consistent format, and the survey structure helped capture both instinctive reactions and more thoughtful reflections. With all the data cleaned, sorted, and grouped, I was ready to move on to the next stage of the project.

Evaluation

Introduction to Evaluation

This section looks at the results of the survey and will show how well people responded to the real and deepfake videos. The aim of the survey was to see how accurate participants were overall, how confident they felt about their decisions, and what kinds of clues they used to make those choices. By combining the multiple-choice answers with the question that required text it made it easier for me to get a deeper understanding as to why the people completing the survey thought the video was real or fake, what gave it away to them and also a freedom for them to voice their own opinion on the videos. This also helped show where people were guessing correctly and where they were being misled which is really important when thinking about how this technology might affect everyday media use.

Accuracy

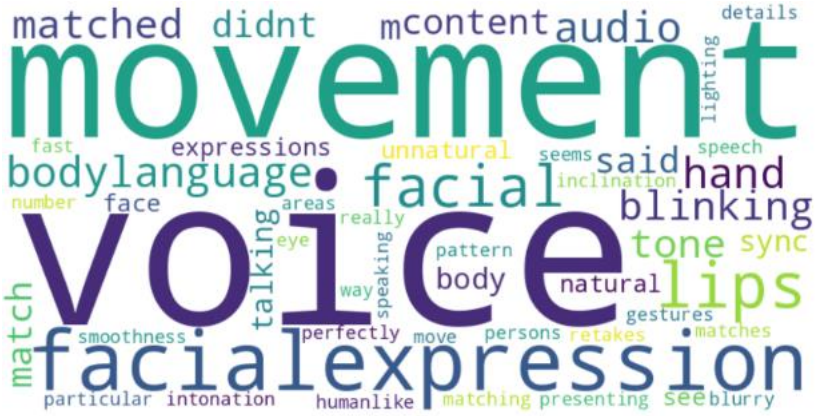
The results showed that participants were generally able to identify deepfakes with reasonable accuracy. For example, Videos 3, 5, and 6 which were deepfakes were correctly labelled by most participants, with 13 to 15 responses identifying them as fake. This shows that people do pick up on subtle signs when something feels artificial. However, there was also a significant number of false positives, especially in Videos 1, 2, and 4, which were real but still flagged as fake by several people. In Video 2 alone, five participants believed it was a deepfake, even though it was authentic. This suggests a growing uncertainty or even over

cautiousness, where people might label something as fake just in case. The results of the survey are in a table below.

Video Number	Classified as Deepfake	Classified as Real	Selected 'Not Sure'	Actual Video Type
Video 1	2	14	1	Real
Video 2	5	11	1	Real
Video 3	13	2	2	Deepfake
Video 4	2	13	2	Real
Video 5	15	1	1	Deepfake
Video 6	13	5	1	Deepfake

Confidence levels were also quite high overall, with most people reporting they felt “Confident” or “Very Confident” in their answers even in cases where they were incorrect. This points to a mismatch between how sure people feel and how accurate they are which could be risky when engaging with misleading media online. As shown in the word clouds below participants relied on visual and behavioural clues like lip syncing, facial expression, eye movement and voice tone to make their decisions. These terms appeared most frequently across both real and fake videos, showing that while people are clearly looking for patterns, those patterns don’t always align with reality. It is something to note in the future that these were the key signs that most participants looked out for. Hopefully down the line tools or training courses this will be beneficial.

Real Video Work Cloud

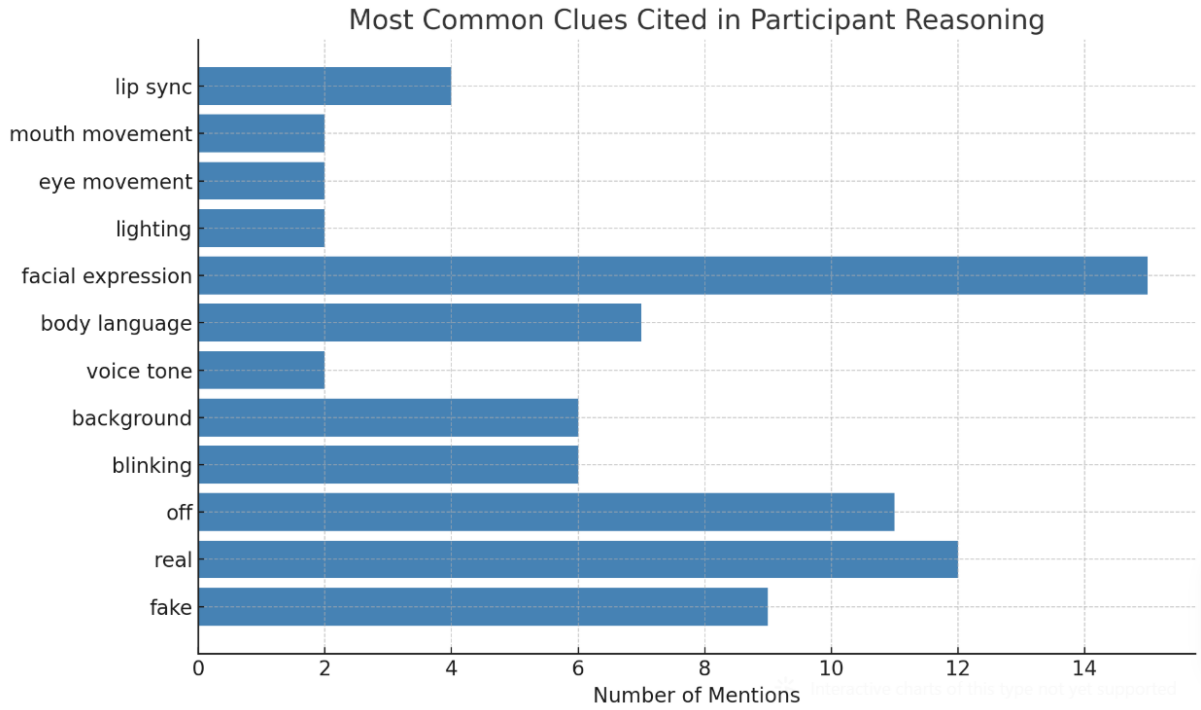


Fake video word cloud.

Individual Observations and Overall Impressions

Beyond the general patterns, a few individual responses offered interesting insight into how people approached the task. One participant pointed out inconsistencies in background lighting and how it didn't match the speaker's positioning something that wasn't commonly noticed by others. Another highlighted mouth movement not aligning with speech but still believed the video was real because the subject's body language looked natural. Below is an image made from the specific signs participants noticed after each video and the number next to them represents how common those specific signs were mentioned in my survey. There were also cases where participants said they recognised the person or had seen the video before which clearly influenced their decision. This shows that recognition and memory sometimes played a role not just visual analysis.

Across all responses there was a clear effort to give reasoning even when vague and most people attempted to justify their choices rather than guessing blindly. Some responses were very thoughtful while others admitted uncertainty or relied on instinct. Taken together these results show that people are actively trying to engage with the concept of deepfakes, but many are still unsure of what to trust. The mix of observation, emotion and guesswork suggests that while awareness is growing the ability to detect deepfakes consistently remains a work in progress.



Summary

Overall the results from this study confirm that while participants are increasingly aware of deepfakes, their ability to detect them still varies widely. Most people relied on visual cues and instinct rather than technical knowledge which would to be expected and while some responses were thoughtful and observant others were more of a guess. Interestingly, even when participants misidentified videos, they often felt confident in their choices, pointing to a growing challenge in trust and perception in the digital space. The patterns in reasoning, supported by the word clouds and frequency analysis, show that certain clues are being repeatedly used such as mouth movement and facial expression but these aren't always enough to guarantee an accurate judgement. These findings highlight the need for better support tools as well as further research into how people interact with synthetic media in everyday settings.

Conclusion

The main goal of this project for me was to see how well a normal everyday person could detect a deepfake video when compared to real videos without any training. Deepfakes are becoming more common and are definitely becoming more realistic as a result it is important to see how the public respond to these different types of videos. Another major reason is the fact that with social media these fake videos are spread faster spreading misinformation. This study focused on the normal humans, Ai systems that detect deepfake content were not part of this survey along with experts on this certain topic. This was a test to see if people could rely on visual clues or else just going off their gut decision to tell what was real and what was fake. The original hypothesis I suggested was that most people would struggle to identify deepfakes consistently and that even when they were confident in their answers, their reasoning might not always align with the video itself.

The responses show some clear trends. Most participants were able to recognise that the deepfake videos were in fact fake saying that there were a couple of instances where a couple of participants thought the videos were 100 percent real. A surprising number of participants also believed that some of the real videos were fake this showed that while people are aware of the concept of deepfakes there is also a hint of uncertainty amongst the participants. Many seemed unsure of the videos as a whole and answered with deepfake just to be safe I would assume. Some of the most common reasons for each individual answer for a deepfake and also the real videos was the facial expressions of the character in the video itself. The lip syncing or mouth movement was also highlighted massively in the results of the participants. These specific signs sometimes pointed to a deepfake but on other occasions it didn't. This shows that people are picking up on certain signs, but those cues aren't always reliable. It also highlights a gap between confidence and accuracy, where people might be guessing or overcorrecting rather than trusting what they see.

These results highlight an important issue people are after coming to terms with not everything you see online is true. They are becoming more cautious about what they see online, but that caution that is evident from the survey doesn't always translate into accuracy. Viewers might be able to tell that what they are seeing is a deepfake but in most instances they don't have the confidence or digital tools to tell 100 percent whether it is real or fake. This leads to second guessing everything. This kind of uncertainty can affect the trust of the people with viewing online content and how they share online content. As deepfake technology continues to go from strength to strength it will also become along harder to detect. This will make it harder for the human eye and will add to the confusion that is already there. The study shows that human judgment alone isn't always enough especially when the differences between real and fake are so subtle although a lot of the right answers were given in the survey the confidence levels weren't at 100 percent also the fact that users knew that they were looking for deepfakes may have altered the study itself. It also suggests that with the right training or awareness, people could become more confident and accurate over time as they already are on the right direction.

In the end this project shows that while people are becoming more aware of deepfakes there's still a lot of uncertainty around how to spot them confidently. A couple of participants could tell when something looked off, but they also doubted the real videos which shows how much trust in video has been shaken. As deepfakes become more advanced we can't just rely on instinct alone. This highlights the need for further research, more public awareness, and possibly even training tools to help people learn and grow from . The next sections of this report will look at some recommendations based on these findings and suggest directions for future work that could help strengthen our response to synthetic media.

Recommendations

One of the clearest takeaways from this project is that more needs to be done to help people feel confident and capable when it comes to spotting deepfakes. While many participants had

a general sense of what to look for, there was still a lot of doubt, especially around real videos that just felt off. This shows that people aren't always wrong they're just unsure. There should be more focus on public awareness campaigns and digital literacy programmes that teach people what deepfakes are, how they work, and what kinds of subtle signs might give them away. This is very easily implemented into the everyday life make it a necessity for people doing these tests prior to starting a job and for younger people in 4th year a suitable time to teach kids about these problems. Even a basic understanding of things like lip-sync issues, unnatural eye movement, or facial glitches could help people become more informed and less easily misled. These are the main 3 signs that differentiate a real video from a deepfake video so even highlighting these and showing comparisons between the two could go a long way.

Another useful step to help and make as many people as possible aware of deepfakes would be to create a simple accessible training tools that let people practise spotting deepfakes in a safe and guided way. This could be something as straightforward as an interactive quiz or game that shows a mix of real and fake videos and gives instant feedback on the viewer's choice. This would be a short and beneficial tool all to promote the safety of online threats. Tools like this could help people sharpen their ability to notice small but important cues especially as the technology keeps evolving. Some versions of this already exist in academic settings or for professional use but they could be made accessible to the public in order to educate them more. Ideally they'd be easy to use on mobile or social platforms where most deepfakes are seen. If people can train their eye and build confidence in a low pressure setting, they'll be better prepared to make informed decisions in real situations.

Social media platforms also need to take a more active role in tackling the spread of deepfakes. Since most people come across this kind of content while scrolling through apps like TikTok, Instagram, or YouTube these platforms have a responsibility to label or flag synthetic media when it's detected especially if it's misleading or designed to deceive. After conducting this Survey I noticed Tik Tok had added captions at the bottom of Ai generated content warning the users that the above video had been made by Ai. Clear labelling helps users approach videos with more caution and can reduce the spread of misinformation. At the same time governments and policy makers should support this by encouraging

transparency rules and developing up-to-date media regulations. This doesn't mean banning deepfakes entirely they have creative uses too for example the media and film industry, but it does mean putting systems in place to protect people from being misled or manipulated. A combined effort between platforms, educators, and policy makers could help slow the spread of harmful deepfakes while giving the public more tools to handle what they see online.

Overall the results of this project suggest that while people are starting to recognise the risks of deepfakes but they still need more support when it comes to detecting them. Improving public awareness, developing simple training tools, and encouraging platforms to be more transparent are all realistic steps that could make a real difference. These recommendations aren't about solving the problem completely deepfakes are always going to evolve but they're about giving people a better chance at recognising what's real and what's not.

Future Work/Changes to the Project

If this project were to be continued or repeated, one of the first things I would do is increase the number of participants and try to include a wider range of people. The Survey I conducted is named a Pilot study and for this survey I relied mostly on friends, classmates, and family, which was fine for a small-scale study but it meant the sample wasn't very balanced in terms of age, background, or digital habits. Getting more responses from people of different age groups, professions, and levels of tech experience could give a much clearer picture of how different groups respond to deepfakes. It might also help show whether certain groups are more vulnerable to being misled or whether others are naturally better at spotting subtle clues. In other words I would just expand the study to more people with different backgrounds and more people in general to get more data.

Another way this project could be developed further is by testing different types of deepfakes or experimenting with a wider range of content styles. In this study the deepfakes were straightforward short videos with a single speaker but there's a lot more that could be explored. This was because of me making the deepfakes myself with an app. There are a lot more realistic and better deepfakes out there, but I wanted to make my own deepfake videos.

For example, future versions could include intense videos, group conversations, or even manipulated audio only clips to see if certain types of videos are more believable then others to the naked eye. It would also be interesting to test how things like background noise, lighting, or editing style influence how people judge a video. As deepfake technology becomes more advanced and more realistic these small details might play a bigger role in how people interpret what they're watching.

It could also be beneficial to distinguish the difference between human responses and Deepfake detection tools in the future version of this project. There is an opportunity to carry out the same project for the people's point of view but just add in the deepfake detection tools and to compare the two different sets of answers. That kind of comparison could highlight the strengths and weaknesses one side might have over the other for example, where AI can spot a glitch, but a person can't, or where human viewers notice something that algorithms miss. Another interesting direction would be to explore group-based detection, where multiple people view and rate the same videos. Some research has shown that crowdsourcing can lead to higher accuracy than individuals working alone, so it might be worth seeing if this really is the case by comparing the two like we will do with the deepfake detection tools. (Somorary ,2024)

As deepfake technology keeps developing the need for ongoing research in this area will only grow and will need to continue developing with the deepfake technology itself. While a lot of attention is given to technical solutions, it's just as important to understand how people interact with this kind of content in real-life settings. Projects like this one can help reveal where people struggle, what clues they rely on and how they make decisions when faced with something that might not be real. Building on this kind of work could help improve media literacy, shape future detection tools and guide how we educate the public about misinformation. There's still a lot to learn but by continuing to explore both the human and technological sides of the problem we can build a stronger response to the challenges that deepfakes present.

References

Ajder, H. P. (2019). *The State of Deepfakes*. Deeptrace.

Alanazi, S. A., & Asif, S. (2023). Understanding deepfakes: A comprehensive analysis of creation, generation, and detection. *Proceedings of the 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2023)*.

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

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).

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39–52.

Kaate, M., Luger, E. and Colley, J., 2023. *Seeing is believing? Exploring user perceptions of deepfake-generated personas in HCI*.New York: ACM.

Kietzmann, J., Lee, L.W., McCarthy, I.P., & Kietzmann, T.C. (2020). Deepfakes: Trick or treat? *Business Horizons*, 63(2),

Groh, M., Shih, M., Turner, S., & Bernstein, M.S. (2021). *Deepfake detection by human crowds,* machines*, and human–machine collaborations*. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1),

Somoray, Klaire & Miller, Dan & Holmes, Mary. (2024). Human performance in deepfake detection: A systematic review. 10.2139/ssrn.4955104.