Journal of Cybersecurity Education, Research and Practice

Volume 2024 | Number 1 Article 1

October 2023

What You See Is Not What You Know: Studying Deception in Deepfake Video Manipulation

Cathryn Allen
Kennesaw State University, calle209@students.kennesaw.edu

Bryson R. Payne
University of North Georgia, bryson.payne@ung.edu

Tamirat Abegaz
University of North Georgia, tamirat.abegaz@ung.edu

Chuck Robertson
University of North Georgia, Chuck.Robertson@ung.edu

Recommended Citation

Allen, Cathryn; Payne, Bryson R.; Abegaz, Tamirat; and Robertson, Chuck (2023) "What You See Is Not What You Know: Studying Deception in Deepfake Video Manipulation," *Journal of Cybersecurity Education, Research and Practice*: Vol. 2024: No. 1, Article 1.
DOI: 10.32727/8.2023.25
Available at: https://digitalcommons.kennesaw.edu/jcerp/vol2024/iss1/1

What You See Is Not What You Know: Studying Deception in Deepfake Video Manipulation

Abstract

Research indicates that deceitful videos tend to spread rapidly online and influence people's opinions and ideas. Because of this, video misinformation via deepfake video manipulation poses a significant online threat. This study aims to discover what factors can influence viewers' capability to distinguish deepfake videos from genuine video footage. This work focuses on exploring deepfake videos' potential use for deception and misinformation by exploring people's ability to determine whether videos are deepfakes in a survey consisting of deepfake videos and original unedited videos. The participants viewed a set of four videos and were asked to judge whether the videos shown were deepfakes or originals. The survey varied the familiarity that the viewers had with the subjects of the videos. Also, the number of videos shown at one time was manipulated. This survey showed that familiarity with subjects has a statistically significant impact on how well people can determine a deepfake. Notably, however, almost two-thirds of study participants (102 out of 154, or 66.23%) were unable to correctly identify a sequence of just four videos as either genuine or deepfake. This study provides insights into possible considerations for countering disinformation and deception resulting from the misuse of deepfakes.

Keywords

deepfakes, disinformation, misinformation, deception, social media, artificial intelligence, image processing, video manipulation

Cover Page Footnote

Research funded in part by Department of Defense Cyber Scholarship grant #H98230-21-1-0197.

What You See Is Not What You Know: Studying Deception in Deepfake Video Manipulation

Cathryn Allen
Institute for Cybersecurity Workforce Development
Kennesaw State University
Kennesaw, GA, USA
calle209@students.kennesaw.edu
0000-0001-5552-968X

Bryson R. Payne
Department of Computer Science and Information Systems
University of North Georgia
Dahlonega, GA, USA
bryson.payne@ung.edu
0000-0003-4539-0308

Tamirat T. Abegaz
Department of Computer Science and Information Systems
University of North Georgia
Dahlonega, GA, USA
tamirat.abegaz@ung.edu
0000-0003-1263-8469

Chuck Robertson
Department of Psychological Science
University of North Georgia
Dahlonega, GA, USA
chuck.robertson@ung.edu
0000-0003-0476-9119

***Abstract*—Research indicates that deceitful videos tend to spread rapidly online and influence people's opinions and ideas. Because of this, video misinformation via deepfake video manipulation poses a significant online threat. This study aims to discover what factors can influence viewers' capability to distinguish deepfake videos from genuine video footage. This work focuses on exploring deepfake videos' potential use for deception and misinformation by exploring people's ability to determine whether videos are deepfakes in a survey consisting of deepfake videos and original unedited videos. The participants viewed a set of four videos and were asked to judge whether the videos shown were deepfakes or originals. The survey varied the familiarity that the viewers had with the subjects of the videos. Also, the number of videos shown at one time was manipulated. This survey showed that familiarity with subjects has a statistically significant impact on how well people can determine a deepfake. Notably, however, almost two-thirds of study participants (102 out of 154, or 66.23%) were unable to correctly identify a sequence of just four videos as either genuine or deepfake. This study provides insights into possible considerations for countering disinformation and deception resulting from the misuse of deepfakes.**

Keywords—deepfakes, disinformation, misinformation, deception, social media, artificial intelligence, image processing, video manipulation.

I. Introduction

Social media is becoming a large part of people's lives. What is posted online is quickly trusted by those who view it. This is due to the many psychological processes that cause people to accept information as truth. Many studies focus on the idea of how people believe false information. Three main ideas that stem from these studies are as follows: automatic belief, cognitive heuristics, and resistance to corrections.

Automated belief theory states that people tend to accept everything as the truth upon viewing it before comparing it to known knowledge [1]. This happens instantaneously even if the information presented is labeled as false [2]. Once people have accepted information as true, it is almost impossible to convince them it is false. Confirmation bias, the interpretation of new evidence as confirmation of a belief, and motivated reasoning, the biased reasoning to produce justifications for one's beliefs, are the root causes of this persistent condition. Motivated reasoning can have an explosive effect on false information because people become passionate about what they believe and will attempt to justify why they think it is true even after being told it is false [1].

The process of comparing is done through cognitive heuristics, which are mental shortcuts to help people solve problems and make decisions quickly. Sometimes cognitive heuristics can make something false feel as if it were true information. One of these ways is through a repetition effect. Anderson [1] mentions that once information has been presented, the next time the information is recalled, it can be associated with the wrong origin and thus convince the brain that the information is true. People subconsciously diagnose the origin of their memories through what details they can recall, the vividness of the memory, and the familiarity they have with the memory [3]. If their brain recalls a memory improperly, it can result in them thinking the information is true.

Resistance to corrections is the phenomenon in which false information has been presented, making it exceedingly difficult to debunk the information. There are two ways in which withdrawing the information can fail: belief persistence and belief echoes [4]. The act of one maintaining their belief despite being shown evidence against it is belief persistence. Motivated reasoning and confirmation bias have turned humans into natural debaters. Motivated reasoning drives people to continue to resist the change in information, while confirmation bias pushes people to find information that supports their claims and ideas without even acknowledging the opposing viewpoint [1]. This keeps individuals closed-minded and not willing to adjust what they know to be true. Sometimes people will correct the misbelief that they have, but there can be a lingering attitude that persists. This is what is called belief echoes [4]. Thorson shows how belief echoes exist after participants were shown false information and then an immediate correction. Those who participated in the survey created an instant opinion once they were introduced to false information. They were then told information that debunked the original claim. It was verified that the participants understood which information was truthful. However, the initial opinion that was formed in the beginning was still present in the end. This survey demonstrates how belief

Research funded in part by Department of Defense Cyber Scholarship grant #H98230-21-1-0197.

echoes continue to influence an individual even when their initial reaction is known to be false.

Deepfakes are videos that have been altered from their original form by artificial intelligence capable of swapping faces, changing audio, or making other changes that manipulate the video's content. Deepfakes were created by a researcher with the tools provided by DeepFaceLab, an open-source deepfake software tool created by Perov, Chervoniy, et al. [5] and posted on GitHub. Figure 1 shows the capability of deepfakes to convincingly replace an actor's face with that of a celebrity or other target. Deepfakes can range from being lighthearted to being deceitful and containing misinformation, information that is intended to deceive. These videos are known to circulate online through social media, and they can be quite deceiving, causing people to believe they are real. Previous research on deepfakes focuses on three factors: the degree of realism in deepfakes, the impact of misinformation and disinformation, and social media as a wildfire or catalyst.

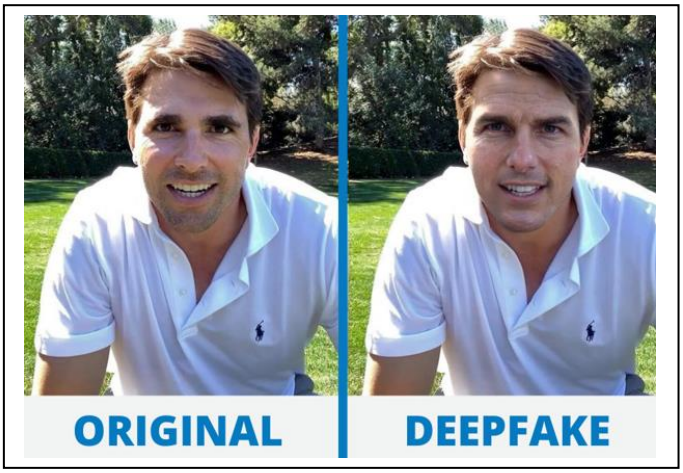

Fig. 1. A sample of manipulated media involving a familiar celebrity, Tom Cruise.

The objective of this research study was to investigate the factors that can influence viewers' ability to determine deepfake videos. Participants were recruited from a university environment including students, staff, and faculty, and they were tasked with identifying which of the four videos shown to them were deepfakes or original videos. The first research goal was to determine the impact of subject familiarity on deepfake detection accuracy by presenting videos of both known and unknown individuals. Familiarity with a subject could include being familiar with one's body language, voice, and speech patterns, including typical vocabulary and pacing. The second goal was to examine the effect of the number of videos presented simultaneously on deepfake detection accuracy. The study examined whether subject familiarity impacted deepfake recognition and whether presenting two videos (one original and one deepfake) side-by-side versus one video (either original or deepfake) affected the determinability of a deepfake.

II. Related Work

A. *The Degree of Realism in Deepfakes*

As discussed in the introduction, deepfakes are typically videos that have been manipulated to contain disinformation by using a variety of different artificial neural networks to replace one person's face with another [6]. People are more likely to be deceived through visuals because they can elicit false memories [3]. Specifically, human memory can be unreliable, eliciting false memories through imagination inflation, which boosts one's confidence that an event took place. This makes deepfakes potentially even more powerful and deceptive than textual information.

Realist-looking deepfakes can be created using a large variety of networks, but typically they are created using generative networks and encoder-decoder networks [6]. This consists of having an encoder and a decoder counter working to produce a realistic deepfake. Deepfake artificial intelligence functions by pitting networks against each other [7]. The process starts with a generator, an encoder, which makes a product based on the input that is given. This is then shared with a discriminator, a decoder that is trying to figure out if the image is real or not. The generator learns new ways to improve its products, thus leaving the discriminator to struggle to tell what is real or fake. This process pins the generator and discriminator against each other causing them to find ways to "outsmart" the other. Through this entire process, artificial intelligence is learning, which is creating deepfakes that become increasingly convincing [7]. Deepfakes, in nature, are very deceiving, and with continuous improvement, they may one day become impossible to detect as deception.

In one study, it was found that only about 50% of the time can people know if a video is a deepfake or if it is real, which is about as good as someone guessing [8]. This suggests that viewers cannot distinguish most deepfakes as fact or fiction. When people cannot tell if a video is a deepfake, the viewers develop a sense of uncertainty. Chadwick and Vaccari revealed that 35.1% of viewers who watched a 4-second deceptive clip were uncertain if it took place or not, whereas a 26-second video revealed 36.9% viewer uncertainty. This uncertainty causes people to not know what they can trust. Those who express uncertainty about the deepfakes also present with lower trust in social media [8]. This can be a result of deepfakes causing people to question untampered images and videos, which can lead people to no longer believe what they see [9].

B. *The Impact of Misinformation and Disinformation*

When false information is spread, it is called misinformation or disinformation. It is classified as misinformation when the party spreading false information is unaware it is false and classified as disinformation when the party knows that it is false [10]. Deepfakes are categorized as disinformation when they are created and first published and as misinformation when parties spread this false information [11].

Deepfakes are one example of how misinformation and disinformation spread online. These videos can be based upon something ranging from lighthearted and comical to intense and destructive. More than 85% of deepfakes posted online target a woman [12]. Some of these cases show women celebrities, in a pornographic scene. Kristen Bell, Scarlett Johansson, and Taylor Swift have all been victims of these types of deepfakes. These ladies and others not only suffer the emotional toll from the video itself, but they can also face harassment, backlash, and career impacts. In addition to cyberbullying, deepfakes or

altered videos have been known to spread false information about politics. Political deepfakes can create a sense of uncertainty in viewers. This uncertainty can cause viewers to lower their trust in media and may eventually influence their decisions [8]. The spreading of misinformation and disinformation, no matter the media, can be immensely powerful and persuasive.

A video of Nancy Pelosi containing misinformation traversed the internet in 2019. In the video, she was portrayed through altered footage as allegedly drunk by slowing the framerate of the video and slurring her speech. This video has been viewed several million times. Some people who reposted or shared and responded to the video referred to Pelosi as "drunk" and a "babbling mess." Even after this video had been marked as misinformation and debunked, there were still people who were calling for Pelosi's resignation. The emotions that the viewers experienced while witnessing someone in her position slurring words and acting drunk stuck with them even when confronted with the truth that the video was fake. This altered, deceptive video gave viewers reasons to turn against a government official. Previous work indicated that familiarity increases the perception of accuracy when determining if false information is real [13]. This altered video of a high-ranking US government official, third in succession to the presidency, helped spur the research in this paper. Specifically, the research team wondered whether familiarity with a person's body language, voice, and speech patterns could help a typical viewer discern that video of that person had been tampered with, or if seeing the original, unedited video side-by-side with the altered version would be more effective.

C. Social Media as a "Wildfire"

Social media allows for information to travel and spread within seconds. Once posted, there are little to no measures to put out the fire caused by the information. Most people in America receive their news through social media [9]. The world has become heavily dependent on the internet to provide information, entertainment, and resources. This increased dependency becomes dangerous with how information gets published and shared online. Relying on the internet can cause people to trust by default what they read and see online. This even momentary trust can be disastrous when mixed with misinformation that can be amplified, accelerated, and multiplied online with social media as the catalyst.

Social media, as well as the internet overall, is set up in a way that allows misinformation, in multiple forms, to be spread quickly [1]. Social media sites and apps often have little to no limitation on how people can create or purchase accounts. Users can typically create a new account within minutes for free and they are not limited to how many they can create. If an email is provided, the account is created. These emails do not have to be confirmed nor do they need to exist. There are also ways for accounts with high followings to be purchased for a low price. In addition, there are services such as bots that can be purchased that will upload and post content all by themselves. Just like deepfakes, social media bots are not transparent, readily identifiable, or well-understood. These bots can contain artificial intelligence that allows them to write posts that sound like they were written by a human. Over time these bots will continue to improve and get smarter, creating an even more lifelike appearance. People can easily fall for these bots and believe that they are real humans due both to a lack of knowledge and to the bias of automatic belief. The resources that social media offers allow for misinformation to be posted and spread easily, instantly, globally, and permanently.

The rate at which deepfakes can be posted is quicker than the time it takes social media app moderators to review or moderate what is being uploaded. This means that false information can be shared globally before it is ever fact-checked [1]. Even once a deepfake has been fact-checked and removed from the source that posted it, it will often already have been shared or reposted by other accounts. This means that even after deleting the original post, there are other places where the misinformation can be found and accessed. Once a deepfake has been posted online and copied, shared, or re-posted, it becomes virtually impossible to remove it completely from the internet [7].

The World Economic Forum recognized digital misinformation as one of the main challenges to societies globally [14]. Deepfakes are one of the easiest ways to spread misinformation online due to how they manipulate human susceptibility to information. People are naturally drawn to believe what they see, but this could change with the rapid growth of deepfakes. It has been found that deepfakes have begun causing a sense of uncertainty, leaving people to wonder if they can believe what they see online. This can lead to a decline in people's trust in social media. If deepfakes continue to improve at their present rate, they could soon become seamless and undetectable even to expert viewers. With the rapid progression of technology, it is important to analyze deepfakes further.

III. Methodology

The purpose of this survey is to investigate whether specific factors may assist viewers in identifying deepfake videos. Based on the literature reviewed, the survey proposes that subject familiarity and the quantity of videos presented may be key factors to assist typical viewers in determining whether a video is a deepfake. The authors have introduced four hypotheses to test their theory: 1) if the viewer is familiar with the subjects depicted in the deepfake video, they will have a higher chance of identifying it as fake, 2) if the viewer is not familiar with the subjects in the deepfake video, they are more likely to be deceived into believing the deepfake video is real, 3) if viewers are shown two videos, one original and one deepfake, they are more likely to accurately determine which video is the deepfake, and 4) if the viewers are only shown one video at a time, they are more likely to incorrectly identify it as a deepfake, regardless of whether it is an original or deepfake video. By formulating these hypotheses, the authors are attempting to identify key factors that may influence a viewer's ability to detect deepfakes. This research could have significant implications for media literacy and help to inform efforts to combat the spread of disinformation.

A. Survey Structure of Previous Studies

Some previous studies on deepfakes used already-released deepfake videos for their content. Groh et al. [15] and Köbis et al. [16] both used deepfakes found on the Deepfake Detection

Challenge (DFDC). This allowed them access to countless deepfake videos already created. However, doing this posed an issue of all the subjects of the videos being someone that the viewer does not know. So, this new survey incorporates both celebrities and common people to determine if there is a relationship between the familiarity of the subject and being able to determine a deepfake. Chadwick and Vaccari [8] used deepfakes of celebrities that had already been spread online within their survey. Some of the participants in the survey may have already seen the videos, which could alter the results. Because of this, the deepfakes of the celebrities used in this survey were created specifically for this survey.

Groh et al. [15] incorporated both single videos and paired videos for their questions. This means that in one case participants view two videos and pick which one is the deepfake and in the other, they are presented with just one video and must decide if it is a deepfake. This structure allows for a better analysis of participants' understanding of deepfakes. This survey followed the structure of Groh et al. However, Groh et al. decided to have each participant see an equal amount of deepfakes to original videos. At the beginning of their survey, the participants were told that 50% of the videos would be deepfakes and 50% would be original. Revealing that information allowed the participants to make assumptions about what the video may be based on previous guesses. The survey in the current research does not follow the 50/50 ratio, but rather, it has different ratios of deepfakes and original videos per survey.

In their study, Köbis et al. recognized the importance of understanding the participants' confidence levels when answering questions related to deepfakes. By asking for confidence levels, the researchers could differentiate between participants who were lucky guessers and those who were truly able to determine whether a video was a deepfake or not. In a similar vein, in our survey, we also recognized the importance of measuring participants' confidence levels. We utilized a Likert scale, which is a commonly used tool in survey research that measures attitudes and beliefs. The Likert scale used in our survey ranged from 0 to 100, with 0 indicating no confidence at all and 100 indicating complete confidence. After each set of questions, participants were asked to rate their confidence level in their answers. This allowed us to not only analyze the accuracy of their answers but also the level of confidence they had in their abilities to distinguish deepfake videos from real ones.

B. Collection and Creation of Videos

Videos were collected for familiar and unfamiliar cases. For familiar cases, videos of celebrities were found online. For unfamiliar cases, videos were recorded of fellow college students. Some of these original videos were used to create deepfake videos. The deepfake videos were created by using software called DeepFaceLab, written by Perov, Chervoniy, et al. [5]. DeepFaceLab (DFL) is an integrated open-source system that allows for the creation of deepfakes. DFL is structured as a pipeline, allowing for a functioning workflow while also having variations within it.

The first step in the workflow is extraction, the phase in which all the faces from the source and destination videos are taken out and saved. The first step to extraction is face detection for both videos. DFL uses S3FD, which stands for Single Shot Scale-Invariant Face Detector. S3FD is based upon an anchor-based detection framework and has incorporated a wide range of anchor-associated layers. This wide range of layers sets S3FD apart from other facial detection frameworks because it allows for facial detection to happen no matter how small or large the face is within a video [17]. After facial detection, facial alignment takes place. This is the process of determining facial landmarks, which can consist of the edges of the eyes, the edges of the mouth, the bottom of the nose, and other facial features. These facial landmarks allow for facial stability between the two faces by being able to pinpoint where features need to be. This takes place by using 2DFAN and PRNET, which are facial landmarking algorithms. 2DFAN focuses on faces with standard front-facing poses, while PRNET works with faces that are not standard. Examples of nonstandard faces are those poses that are turned or tilted away from the camera. There is an optional function that allows for the manual smoothing of facial landmarks. After facial landmarks are collected and saved within corresponding folders, face segmentation takes place. This is done using TernausNet, a fine-grained Face Segmentation network. This network allows for any facial obstructions to the removed from the face-swapping process; this could be a hand, glasses, a hat, etc. [5].

The second step in the pipeline is training, which introduces the DF (original deepfake autoencoder) and LIAE (lighting-improved autoencoder) structures. The DF structure is based upon an Encoder and Inter (short for Interpolator) that have shared weights between the source and destination and two Decoders that belong to either the source or destination. Although DF can produce a deepfake video, it struggles to inherit enough information to produce a seamless video. The information that is not addressed in DF is light consistency. The LIAE autoencoder structure is more complex than DF because it has a shared-weight Encoder, Decoder, and two independent Inters. The LIAE structure takes the latent code produced from the Inters and concatenates them through a channel to produce a deepfake video that includes light consistency [5]. This means that during the training process, the LIAE structure will maintain consistent lighting between the source and destination videos. This will help create a deepfake that looks seamless and has the same lighting throughout the whole video.

The third and final step in the pipeline is conversion or merging. This is when the face within the destination video is replaced with the desired source face. This is done by taking the desired face and mapping it alongside the original facial landmarks in the video. This will put the desired face in the correct location and with the same facial expression as the original face in the video. After the face is inserted, DFL uses color transfer algorithms and Poisson blending to make the newly added face look seamless and match the correct skin tone, within a reasonable delta. Finally, there is a pre-trained face super-resolution neural network within DFL that sharpens the blended face and does its best to add details back to the bleak and smoothed-over face [5]. Each deepfake required approximately 24-30 hours to complete following this workflow.

C. Structure of the Surveys

The collected original videos and created deepfake videos were used to produce surveys. Each survey had four question sets, each focused on different videos. The question sets differed in the familiarity of the subject in the video and the number of videos displayed at one time. This produced the following sets: 1) unfamiliar subject with one video, 2) unfamiliar subject with two videos, 3) familiar subject with one video, and 4) familiar subject with two videos. Each question set contained questions that followed the videos shown. For any question set that involved a celebrity, a question was asked to determine if the participant knew who the celebrity was. This was done by having a multiple-choice question listing four celebrity names. If the correct name was selected, the data for this question was used as a familiar case and if an incorrect name was selected, the data was ignored.

For single video question sets, the participants were asked a multiple-choice question asking if the video shown was a deepfake or original video. For the question sets with two videos, the participants were asked a multiple-choice question asking which video was the deepfake between the two displayed. The next question asked was a Likert scale to determine the participant's confidence regarding their previous answer. This scale ranged from 0, not confident at all, to 100, being fully confident. The final question asked the participants to explain why they believed the answer they selected. This question was an open text box so participants could write anything they wanted.

The unfamiliar subject cases showed a video of a student, while the familiar cases showed a video of a celebrity. Each survey randomized the order, what subjects were used for each condition, and whether the participants were presented with original videos or deepfakes. The goal of these measures was to help prevent participants from assuming a video was a deepfake or an original.

D. Survey Administration

The survey was virtual and was advertised to faculty, staff, and students at the University of North Georgia, excluding those who are in close contact with or would know any of the people used for the unfamiliar cases. It was an anonymous survey with no incentives provided. Therefore, no personal data or demographics were collected from the participants. The surveys started with an instructional portion to explain and describe deepfakes to the participants. The participants were then given a survey with 4 question sets.

IV. Results

In this study, of all the participants who were recruited, 154 completed the survey. Any data that was not fully complete was thrown out. If a participant did not answer the correct sample video, their data was ignored for the question when familiarity mattered. To determine if the familiarity of subjects had an impact on how effective participants were at determining if a video was a deepfake, the accuracy percentages of unfamiliar (the subjects of the deepfakes are someone that the viewer is unfamiliar with,) and familiar (subjects of the deepfakes are someone that the viewer is familiar with) questions were computed. It is important to note that accuracy is defined as the percentage who identified the deepfake video, not the people who "accurately" identified the video as real.

Overall, almost exactly two-thirds (66.23%) of participants in the study incorrectly identified at least one out of a series of four videos as either authentic or deepfake-generated. Participants were college-educated adults (from 18 to 70-plus years old), in an IRB (institutional review board) approved study, and all were informed that the topic of the study was distinguishing deepfake videos. If just over two-thirds of informed, college-educated individuals were unable to successfully distinguish deepfakes from authentic videos when they were actively looking for them, the implications for deepfake video disinformation and deception in reflexively shared social media posts are potentially dire.

A. Descriptive Statistics of Hypotheses 1 and 2

To determine if familiarity with the subjects had an impact on how adept participants were at determining if a video was a deepfake, the accuracy percentages of unfamiliar and familiar questions were found for when participants were given a deepfake video. Table 1 contains the descriptive statistics for the first two hypotheses. From the information in Table 1, the overall mean percentage of accuracy difference between unfamiliar and familiar videos is 22.46%. This shows that there was a 22.46% increase in accuracy between familiar versus unfamiliar video subjects. A t-test, with an alpha value of .05, was performed to analyze this data in search of significance. The found t-value was 2.209, which is greater than the needed t-value of 1.96 to prove significance. Therefore, there is a 5% significance level that there is a difference between familiar and unfamiliar accuracies.

TABLE I. Percentages of Accuracy for Unfamiliar and Familiar Questions

	Unfamiliar Accuracy	***Familiar Accuracy***
One video at a time	43.9%	72.7%
Two videos side-by-side	82.2%	95.3%

The first focus of this study involved analyzing whether unknown subjects cause deepfakes to be harder to recognize. Since it can be supported that the familiarity of the subject plays a role in how well someone can determine if a video is deepfake, not knowing how someone of power or authority looks, speaks, or behaves, in general, would leave one vulnerable to deepfakes about them. A typical viewer should be educated to know that a lack of familiarity with the subject of a video could leave the viewer susceptible to falling for deepfakes.

B. Descriptive Statistics of Hypotheses 3 and 4

To determine if the number of videos shown at one time had an impact on how well participants could determine if a video was a deepfake, the accuracy percentages of one video and two video question sets were found. From the data in Table 1, the percentage accuracy mapping between one video and two videos clearly shows that using multiple videos improves the accuracy, by at least 10%. The overall accuracy mean difference between using a single video versus using multiple videos is 17.26%. A t-test, with an alpha value of .05, was performed to analyze this

data in search of significance. The found t-value was 1.320, which is less than the needed t-value of 1.96 to prove significance. This accuracy difference failed the t-test for statistical significance, meaning that the presentation of two videos side-by-side (one genuine and one deepfake) did not statistically significantly improve the likelihood that a viewer could successfully identify the deepfake video.

This aspect of the survey focused on whether having two videos presented side-by-side versus one video at a time changes how determinable a deepfake is. Since it was not found that deepfakes presented on their own are more influential than deepfakes presented with other videos, it cannot be supported that the quantity of videos matters. This leaves many questions regarding how to counteract a deepfake. This survey displayed a deepfake and an original video side-by-side in hopes of increasing participants' accuracy. However, there was no statistically significant increase in accuracy—having the original, unedited version of the deepfake video next to the deepfake did not help people determine which video was the deepfake.

C. Perceived Confidence of Participants

After each question set, the participants were asked to provide how confident they felt with their answer on a scale from 0-100. This scale was a Likert scale, so any whole value could be selected. In Table 2 are the mean perceived confidence scores for all participants sorted by the following types: 1) 1 Video Question, 2) 2 Video Questions, 3) Unfamiliar Video Questions, and 4) Familiar Video Questions.

TABLE II. PERCENTAGES OF ACCURACY FOR UNFAMILIAR AND FAMILIAR QUESTIONS

Question Type	Perceived Confidence Mean
1 Video Questions	74.74%
2 Video Questions	70.53%
Unfamiliar Video Questions	65.49%
Familiar Video Questions	78.94%

T-tests, with alpha values of .05, were performed to analyze the confidence of participants further to see if there lies any significance in this data. When comparing one-video questions to two-video questions, the found t value was 1.319, which is less than the needed t value of 1.96. Therefore, there is no significant difference in confidence between one-video and two-video questions. However, when the perceived confidence of unfamiliar questions was compared to that of familiar questions, the found t value was 4.519, which is greater than the need t value of 1.96. Therefore, there is a 5% significance level between the perceived confidence of unfamiliar and familiar questions. This shows that participants had more confidence when answering a question about someone they were more familiar with.

V. CONCLUSIONS AND FUTURE WORK

A. Countering Disinformation and Deception Caused by Deepfakes

The research shows that familiarity with an individual depicted in a deepfake video can influence whether a viewer can determine if the video is altered. However, if the target of a deepfake video is unfamiliar with the subject, not even showing the original and edited videos side-by-side may reliably help a viewer determine the authenticity of either video. This has some important implications for countering deepfakes in practice.

Perhaps most striking is the fact that only 33.77% of participants could correctly identify all four videos as either genuine or deepfakes. Roughly two-thirds (66.23%, or 102 out of 154) of survey participants misclassified at least one video as either deepfake when it was authentic, or potentially worse, they believed one or more deepfake videos to be genuine. The potential for deepfakes to confuse or misinform a majority of the public via social media should not be underestimated.

It would appear, based on the results of this study, that addressing deepfakes could be better achieved by helping a viewer become more familiar with the subject of a deepfake video. Seemingly paradoxically, our results indicate that this approach might be more effective than even showing a deepfake video alongside the original source video(s) used in creating the deepfake. Familiarity with the target individual depicted in a video contributed to viewers' accuracy in distinguishing a deepfake better than showing the authentic source videos side-by-side with the deepfakes.

A contemporary example of this approach appeared during this research, as a deepfake video of Ukrainian President Volodymyr Zelensky surfaced in March 2022, appearing to urge Ukrainian citizens to surrender to Russia's ongoing invasion of the country [18]. President Zelensky's team was prepared for potential Russian disinformation, and they responded in near-real-time to the deception. Within minutes of a television station's mention of the video, President Zelensky posted a Facebook video discrediting the deepfake video and denying the video's message. Facebook, Twitter, and YouTube deleted uploads of the video for violating terms of service related to deceptive and/or manipulated media. It is possible, based on the results of this limited research, that Zelensky's use of a live video of himself was as important as the content of the repudiation message itself.

Organizations, governments, and individuals seeking to contain or counter deepfake deception will need to consider both factors above in their operational planning: 1) a swift, near-real-time response, and 2) creating more familiarity through additional, preferably live video footage of the target of the deepfake responding to and refuting the disinformation personally.

B. Future Work

Overall, this study provides insight into how capable people are at determining if a video is a deepfake or authentic video. The survey was able to support that familiarity with subjects plays a role when determining deepfakes. However, the software used to create these deepfakes did not have very advanced

capabilities or features for editing videos beyond the deepfake face-swapping component. With more advanced software, additional head movements and actions may be incorporated to add to the deception. If possible, adding AI-modified voices for each subject can help create a more realistic video. For future studies, the authors posit that the more realistic the deepfake, the more deceiving it will be.

In academic research, the sample size and selection are critical components that can impact the generalizability and validity of the results. The authors recognize that the sample size of participants used in this study was 154 college-educated adults. It is acknowledged that the results obtained may differ if a larger set of participants were used. The authors have plans to expand the study cohort size and conduct further research on combating deepfake videos. One area of interest involves investigating whether the dissemination of additional unaltered videos of the individual depicted in the deepfake could address the issue of disinformation. This would aim to familiarize viewers with the target individual in question. Another aspect of the future work involves training computer models to better detect deepfakes, by providing similar training with authentic source videos of a target to the AI models.

REFERENCES

[1] K. Anderson. Truth, Lies, and Likes: Why Human Nature Makes Online Misinformation a Serious Threat (And What We Can Do About It). *Law & Psychology Review*, *44*, 2020, 209–243.

[2] J. Donahue and M. Green, Persistence of Belief Change in the Face of Deception: The Effect of Factual Stories Revealed to Be False. *Media Psychology*, *14*(3), 2011, 312–331. https://doi.org/10.1080/15213269.2011.598050

[3] S. J. Frenda, E. D. Knowles, E. F. Loftus, and W. Saletan, False memories of fabricated political events. *Journal of Experimental Social Psychology*, *49*(2), 2013, 280–286. https://doi.org/10.1016/j.jesp.2012.10.013

[4] E. Thorson, Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication*, *33*(3), 2016, 460–480. https://doi.org/10.1080/10584609.2015.1102187

[5] I. Perov, N. Chervoniy, et al. DeepFaceLab: Integrated, Flexible, and Extensible Face-Swapping Framework. 2020. https://arxiv.org/pdf/2005.05535.pdf.

[6] W. Lee and Y. Mirsky, The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, *54*(1), 2021, 1–41. https://doi.org/10.1145/3425780

[7] N. Sanders and J. Wood, Dealing with "Deepfakes": How Synthetic Media Will Distort Reality, Corrupt Data, and Impact Forecasts. *Foresight: The International Journal of Applied Forecasting*, *59*, 2020, 32–37.

[8] A. Chadwick and C. Vaccari, Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*, 2020. https://doi-org /10.1177/2056305120903408

[9] A. McPeak, The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood. *Vanderbilt Journal of Entertainment & Technology Law*, *23*(2), 2021, 433–450.

[10] V. Dan, B. Paris, J. Donovan, M. Hameleers, J. Roozenbeek, S. van der Linden, and C. von Sikorski, Visual Mis- and Disinformation, Social Media, and Democracy. *Journalism & Mass Communication Quarterly*, *98*(3), 2021, 641–664. https://doi.org/10.1177/10776990211035395

[11] T. Dobber, N. Helberger, N. Metoui, D. Trilling, and C. de Vreese, Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *International Journal of Press/Politics*, *26*(1), 2021, 69–91. https://doi.org/10.1177/1940161220944364

[12] S. Brenner, P. Filkuková, J. Langguth, K. Pogorelov, and D.S. Schroeder, Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication*, *6*, 2021. https://doi.org/10.3389/fcomm.2021.632317

[13] G. Pennycook, D.G. Rand, Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 2020, 185–200. https://doi.org/10.1111/jopy.12476

[14] J.C.W. Brooks, U.K.H. Ecker, A. Gordon, S. Lewandowsky, and S. Quadflieg, Exploring the neural substrates of misinformation processing. *Neuropsychologia*, *106*, 2017, 216–224. https://doi.org/10.1016/j.neuropsychologia.2017.10.003

[15] M. Groh, Z. Epstein, C. Firestone, and R. Picard, Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(1), 2022, 1–11. https://doi.org/10.1073/pnas.2110013119

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

[17] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S.Z. Li, S3fd: Single shot scale-invariant face detector. In *Proceedings of the 2017 IEEE international conference on computer vision*, 2017, pp. 192-201. https://doi.org/10.1109/ICCV.2017.30

[18] T. Simonite, A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be. Wired.com, March 17, 2022. Retrieved from https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/