The Report Committee for Lu Jin
Certifies that this is the approved version of the following Report:

Cheated by Deepfakes?
Deepfake Detection Ability, People's Reactions, and Ethical Implications

APPROVED BY
SUPERVISING COMMITTEE:

Kenneth R. Fleischmann, Supervisor

Danna Gurari

Cheated by Deepfakes?
Deepfake Detection Ability, People's Reactions, and Ethical Implications

by

Lu Jin

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Information Studies

The University of Texas at Austin

May 2020

Acknowledgements

I would like to thank Dr. Kenneth R. Fleischmann for serving as my supervisor and Dr. Danna Gurari for serving as my second reader for this interesting and inspiring master's report. I also appreciate the guidance and technical support provided by Nitin Verma. Given the challenges of this unexpected special situation of a pandemic shut-down, completing this report is particularly special and memorable.

Abstract

Cheated by Deepfakes?
Deepfake Detection Ability, People's Reactions, and Ethical Implications

Lu Jin, M.S.INFO.ST

The University of Texas at Austin, 2020

Supervisor: Kenneth R. Fleischmann

Recent dramatic developments in the fields of computer vision and deep learning technology have opened up a range of possibilities not previously imagined. The applications of computer vision technology include manipulating any face in any video and changing the environment of photos, just to name a couple of the new applications. However, these applications are already having impacts on our everyday lives. Given these recent advances in computer vision technology, people may not be able to trust images and videos we see on any media channel. These videos and images have the potential to deceive us.

Throughout the history of technology development, the pros and cons of new technology are often in dispute. New technology is often sensationalized in terms of the benefits for people, which may go beyond anyone's control and imagination. For example, the internet was started with a goal of developing a decentralized network. However, due to how it was commercialized in use, the Internet actually became more centralized than had been intended. Since a centralized platform has the advantage of controlling all

users' data and information, these can be sold to companies to help them engage in targeted marketing. Thus, the Internet fell short of its expectations and hype. Now, the focus and hype has largely shifted to artificial intelligence. In which direction will these new technologies go? What are humans' relationships with these emerging technologies? How can we use this technology safely and ensure that it leads to a future that we want? This goal is the starting point for this report.

In this report, I will use the latest FaceForensics++ dataset as a base for an experiment to answer three research questions: First, how well do people detect deepfakes, and what factors affect their ability to detect deepfakes? Second, what are their reactions when deepfakes are revealed ? Third, what do they see as the ethical implications of deepfakes, and how deepfakes could be used or abused?

For RQ1, I explore the elements that can help people detect deepfakes. For RQ2, I evaluate their reactions. For RQ3, I explore how they perceive the ethical implications of deepfakes. More generally, my findings offer guidance for thinking about how to rebuild trust in video data in an era of deepfakes?

Table of Contents

List of Tables

List of Figures

Introduction

Given the recent dramatic developments in computer vision and deep learning technology, computer vision is, in some cases, starting to match and outpace human vision. The applications of computer vision technology include manipulating any face in any video and changing the environment of photos, among many others. However, these applications will likely result in some dramatic societal impacts. With the computer vision technology increasingly impacting our everyday lives, we, humans, cannot automatically trust images and videos we see on any media channel. These videos and images could potentially cheat us and impair our judgement.

In the history of technology, we can see many pros and cons of new technology. New technology often starts with benefits for humans and goes beyond anyone's control and imagination. For example, the internet started as a decentralized network, and its developers did not intend for it to be centralized. However, due to commercialization, the Internet eventually became highly centralized. Since a centralized platform has the advantage of controlling users' data and information, these valuable commodities can be sold to product companies or brand owners to help them target audiences for services and ads. On the contrary, the decentralized platform is not as easy to commodify. Now, the latest technology - artificial intelligence (AI) - has come to our world. Which direction will these new technologies go? What are humans' relationship with these latest technologies? How can we use technology safely and control the latest technology to shape our future to the way we want it to be? These are the starting points for this paper.

In this paper, I will use the latest FaceForensics++ dataset as the base for an experiment to answer three main questions:

- First, how well do people detect deepfakes, and what factors affect their ability to detect deepfakes?
- Second, what are their reactions when deepfakes are revealed ?
- Third, what do they see as the ethical implications of deepfakes, and how could they could be used or abused?

Related Work

TECHNOLOGY

History of AI and Deepfakes

The starting point of AI can be traced back to the classical philosophers who tried to depict human thinking as a symbolic system. In 1956, a conference at Dartmouth College was where the term "artificial intelligence" was first coined. AI experienced its first "AI winter" recession during the 1974-1980 and then was revived in the 1980s. Another "AI winter" arose from 1987 to 1993. In 1997, IBM's Deep Blue became the first computer to beat a chess champion when it defeated Russian grandmaster Garry Kasparov. With the power of computer hardware, AI has attracted attention again at the beginning of the 21st century since machine learning has been successfully applied in academia and industry.

Computer vision is a subset field of artificial intelligence that trains computers to interpret and understand the visual world. Its goal is to understand the content of digital images. Deep learning is one of the machine learning methods in AI that aims to imitate the mechanisms of the human brain in getting data and forming patterns for use in humans' decision making. It is able to learn in an unsupervised fashion from data that is unstructured or unlabeled. Deep neural networks (DNNs) is a term used interchangeably with deep learning, to capture that the methods involve an artificial neural network with multiple layers between the input and output layers.

The main principle of deepfakes (Güera & Delp, 2018) is to use deep neural networks to replace one face with another face in videos. The technique typically involves auto-encoders or generative adversarial networks (GANs), which use the techniques of an

encoder to reduce an image into a lower dimension and a decoder to reconstruct the image.

FaceForensics++ (Rossler, 2019) is a dataset that contains 1000 original video sequences that have been manipulated with four automated face manipulation methods. It originated from 977 videos collected from YouTube. This dataset contains a dataset of over 3000 manipulated videos from 28 actors in various scenes.

The Future of Deepfakes

Deepfakes are growing in prominence. At present, this technology is not widely used beyond research teams and a few amateur hackers, and the technology is not mature. However, we can foresee that it is the kind of technology that can easily be adopted by the majority of persons. Due to a limited degree of regulation of information technology, it may be hard for the government to ban the masses from using this technology because it does not need any supply chain. Chesney and Citron (2019) express concerns about the adoption of deepfakes. They hold the view that even if deepfakes can bring benefits, they will also execrable the truth decay. It can lead individuals and businesses to be exposed to new forms of exploitation, intimidation, and personal sabotage.

However, Sam Lessin (2019) shows his insight on deepfakes, he argues that the problem relates to the distribution of deepfakes technology. Thus, it will be fairest for this technology to be widely distributed rather than only used by a minority of people. He holds positive attitudes towards the future of deepfakes. He thinks that deepfakes will have a positive effect on human life rather than a negative effect. And, he believes that humanity will come up with new ways to differentiate between real and fake content. The positive effect he mentioned is that we can use the deepfakes to enhance our privacy - which we lack in this highly technology-mediated environment.

From my understanding, everything has pros and cons. We cannot foresee the future until it comes. Deepfake technology is quite like disinformation or virtual reality technology, it seems to provide value but also can trigger negative effects. But if it is as an invention of humans and it has value to some extent, it may not disappear and at the same time it is hard to use power to forbid this invention. In order to face this situation, we need to prepare well before its largely being deployed into the reality and reduce its negative impacts beforehand.

TRUST

Technology Ethics

There is an old saying, "guns don't kill people, only people kill people." The meaning is that since only humans are capable of moral actions, there is no ethical problem with developing and deploying technologies such as guns, or machine vision, only in how end users use these technologies. However, in reality, new technologies enable new human behaviors, so indeed there is a relationship between ethics and technological development. Therefore, it is worthwhile to discuss what humans should do in the face of a new technology. Can we lead the technology to result in a the better situation rather than a worse situation?

With respect to the Internet, you can see it leads to an unwanted situation. Large internet companies dominate the world. User data is automatically collected and used for profit. The news we see is recommended by a machine, which might narrow our mindset and control our minds. It drifts away from its original intentions.

For the development of computer vision and deep learning, now the technology can easily manipulate and create deepfake videos. Therefore, it has become easier than

ever to fabricate someone appearing to say or do something they did not really do. It disrupts our social channels, meaning we cannot trust any images or videos because they can be manipulated intentionally. How can we face this upcoming situation? Some new applications come as a response to the safety of technology. For example, in order to protect and help people to detect fake videos that were created by AI, a researcher has invented a reality defender that can detect the fake content automatically right in the browser. Some laws have been invented to ban the use of deepfakes to produce misinformation - A new California law, which goes into effect next year, will make it illegal to distribute AI-altered audio or video clips that portray politicians in a damaging or demeaning light within 60 days of an election.

The Underlying Principle of Human Trust

Trust is a concept that has been studied in fields such as sociology and psychology. Sociologist Diego Gambetta (2000) explains, "...trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent will perform a particular action, both before [we] can monitor such action (or independently of his capacity of ever be able to monitor it) and in a context in which it affects [our] own action" (Diego, 2000, p. 4).

From history, we know that trust is a tendency in our evolution. We hold trust through our birth, we relied on our mother to feed us with our total trust. We build mutual trust when we build relationships with outsiders. Research shows that brain chemistry governs our emotions and also has an effect on trust. Why do we trust sometimes and not trust at other times?

Kramer (2009) explains that “human beings are naturally predisposed to trust—it’s in our genes and our childhood learning—and by and large it’s a survival mechanism that has served our species well” (Kramer, 2009, p. 3).

We find it easier to trust something that is familiar to us. If we can have physical connection with others, it will build more trust. Many findings indicate that we are more likely to trust others than to distrust others.

There is a lack of research on human trust in deepfakes. As Vaccari and Chadwick (2020) mention, there is a lack of academic research on the effects of deepfakes. They express a concern that deepfakes might “cultivate the assumption for the people that a basic ground of truth cannot be established” (Vaccari & Chadwick, 2020, p. 3). User Emotion

Ortony, Clore, and Collins (1988) discussed 22 human emotion types which are grouped into eight categories. They are well-being, fortunes of others, prospect based, confirmation, attribution, attraction, well-being/attribution and attraction/attribution. In the survey design, open questions are designated to make people express their feeling towards deepfakes. I seek to detect their responses to analyze the transformation in humans’ minds. Will they feel comfortable when they know they cannot detect a fake face? How do these emotions transform and do they lose trust towards videos afterwards?

Research Design

Research Question

My research is driven by three key questions:

• First, how well do people detect deepfakes, and what factors affect their ability to detect deepfakes?

• Second, what are their reactions when deepfakes are revealed, and what is their emotional response?

• Third, what do they see as the ethical implications of deepfakes, and how could deepfakes be used or abused?

For RQ1, we will explore the elements that can help people detect deepfakes. For RQ2, we will evaluate their responses. For RQ3, we will explore how they perceive the ethical implications of deepfakes. More generally, our findings will offer a foundation for thinking about how to rebuild trust in video data in an era of deepfakes.

Data Collection Platform and Corpus

In order to answer my research questions, I designed an experiment using Amazon's Mechanical Turk (MTurk) crowdsourcing platform. The advantages of MTurk (Verma, Fleischmann, & Koltai, 2017, 2018, 2019) are its demographic diversity and accuracy rate of results compared to the survey pool from other platforms or circulating within universities. Ten fake face videos were selected from the Faceforensics++ dataset (Rossler 2019). Another 10 pairs of real face video were selected from YouTube without any famous persons. Each video lasts approximately 15 seconds. Due to the immaturity of the technology, some deepfake videos did not replace the face well. All of the videos are muted to focus on the visual aspect of the videos.

EXPERIMENTAL DESIGN

The experiment begins with the user completing an informed consent form, requiring users to acknowledge that they are 18 years of age or older and that they consent to participate in the research, following the guidelines of the UT-Austin Institutional Review Board (IRB). Next, users fill out a pre-survey, including questions related to their familiarity with and attitudes toward deepfakes, as well as their own self-efficacy in relation to detecting deepfakes. Next, I show the user a randomly generated set of ten faces, sequentially and in randomized order, with random selection of either a real face or a deepfake. After viewing each video, they will be asked if the face was a real video or a deepfake. After that, I asked the questions about their attitudes again, along with questions that ask about their confidence in their ability to successfully complete the exercise and their emotional response. Then, I showed them the ten images, their answers, and the ground truth. I asked them yet again about their attitudes, self-efficacy, and emotional response, and I also asked open-ended questions about the ethical implications of deepfakes and how they view different uses of deepfakes.

The survey was posted as a Human Intelligence Task(HIT) on Amazon's Mechanical Turk (MTurk) platform. We aimed to recruit approximately 150 participants. We devised some control questions to ensure that they were taking the task seriously.

PARTICIPANTS

All of the participants were recruited from MTurk. All were over 18 years old. The time they spent on the each page was recorded to validate their response. In addition, some questions were set into the survey to make sure that participants took the survey carefully. At the end of survey, the participants saw a code to enter into the MTurk to re-

ceive the compensation. This study was approved by IRB and is filed as exempt (2020-01-0135).

In this survey, I collected N=155 participants from MTurk. Using the first validation step - the attention questions - I deleted 15 submissions and so kept results from N=140 participants. With the second validation evaluation I set - completion time for the test, I deleted the participants who finished the task in less than 20 seconds. This left N=92 participants which provided results I deemed valid. Among these participants, 64.1% were male, and 48.9% were 30 to 39 years old.

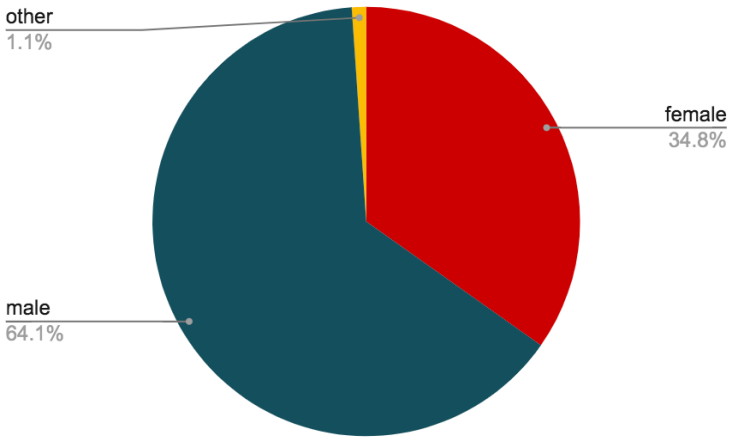

Figure 1. Gender of participants

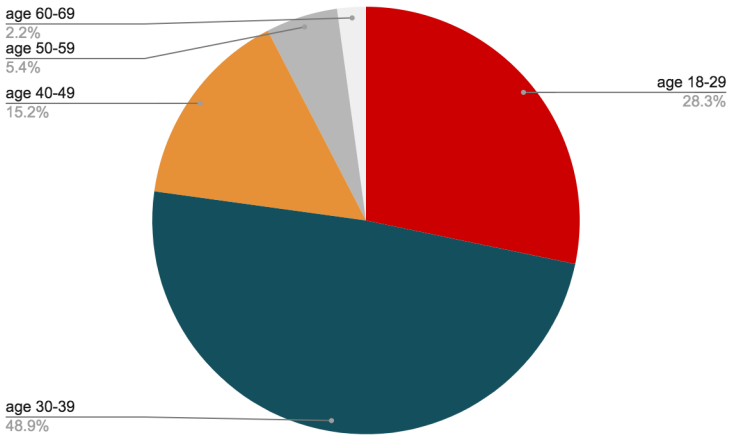

Figure 2. Age range of participants

DATA ANALYSIS

To answer RQ1, I explored the elements that can help people detect deepfakes. I first used some questions to collect whether their familiarity towards deepfakes will help them to detect deepfakes, as well as whether their confidence towards their detection ability had an influence on their ability to detect deepfakes. The most important part is that I tested their detection ability using 10 videos randomized from a corpus that included both deepfakes and real videos. The scores were recorded to show their ability. Descriptive statistics will be used for data analysis on their success rate and the relationship between success rate and age or the familiarity towards deepfakes or confidence towards detection ability.

For RQ2, I evaluated their emotional feedback. After they finished the test, I asked an open-ended question towards how they felt after they completed this question. In addition, after they saw their test score, I asked how did they feel again. From these open-ended questions, I inferred their emotional state at that time.

In RQ3, I explored how they perceive the ethical implications of deepfakes. Will users feel cheated, and what is their emotional feedback? How can we rebuild trust in video data in an era of deepfakes? I used the same question to test participants' concerns about deepfakes before and after testing. This suggests whether their thoughts will change when they felt they lack the ability to detect deepfakes. In addition, I added PVQ questions into this survey to test whether their value system will have an influence on their perception towards deepfakes or the deepfake technology's influence on the future world.

Results

RQ1

How well do people detect deepfakes, and what factors affect their ability to detect deepfakes?

In RQ1, I compared the score with their predicted score, their concerned towards deepfakes, their confidence towards their ability to discern, their knowledge towards deepfakes, and their gender and age. I used *p*-value as a statistical measure to determine whether my hypotheses are correct. If the *p*-value was less than 0.05, then there was a significant relationship between the two variables. In order to get at how well people detect deepfakes and what factors affect their ability to detect deepfakes, I tested the relationship between age and score as well as between gender and score.

From Figure 3, among all of the valid participants (N=92), 17 participants (18.5%) got a perfect score, and 78 participants (84.8%) got a score of higher than 6. From the feedback, they felt more confident when they saw that they got a higher score in the detection test. The results in Figure 4 show the detection score and alongside predicted scores by gender. Females got an average score of 8.03 and males got an average score of 7.36. The results in Figure 5 show the detection score and alongside predicted scores by age. When age is between 18 - 29, the average score is 7.35. When age is between 30 - 39, the average score is 7.84. When age is between 40 - 49, the average score is 7.79. When age is between 50 - 59, the average score is 6.6. When age is between 60 - 69, the average score is 7.

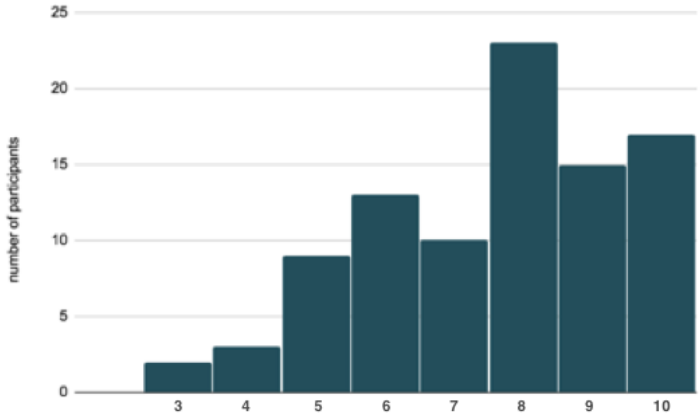

Figure 3. Histogram of detection scores

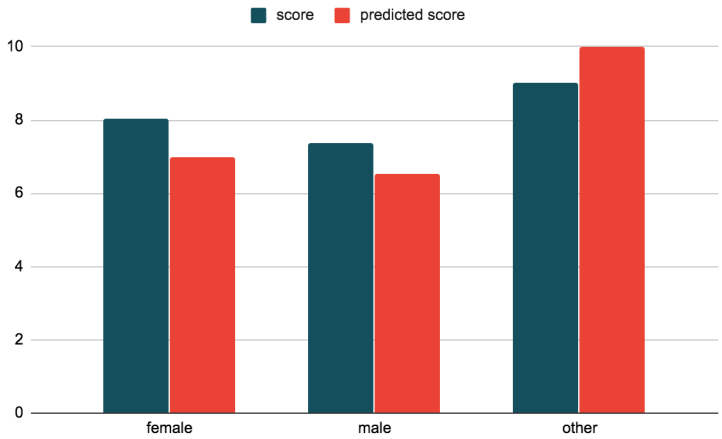

Figure 4. Average detection score and predicted score by gender

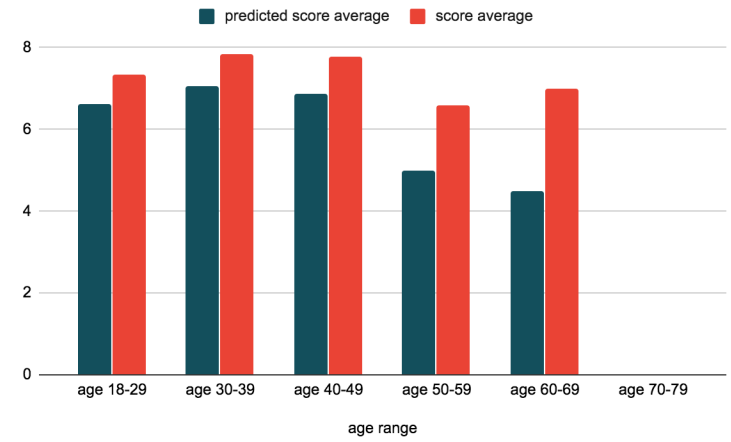

Figure.5 Detection score and predicted score average in age

Figure 6 shows the participants' perceptions towards how accurately they think they can detect deepfakes. It seems most of the participants feel more confident about detecting deepfakes. Their perception of accuracy in detecting deepfakes even increased in the end after the detection test.

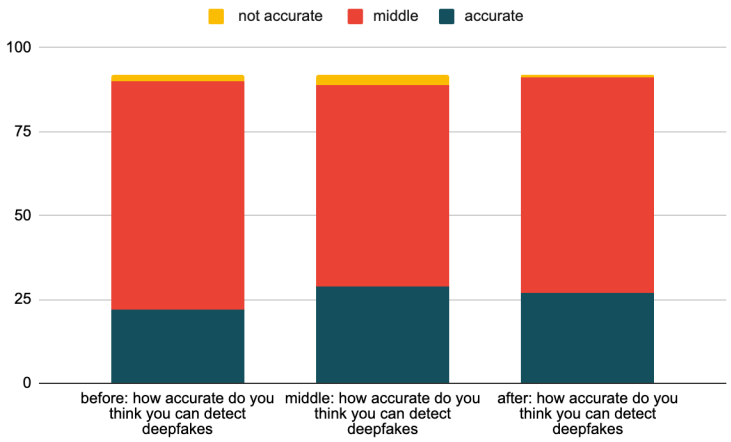

Figure 6. How accurately do you think you can detect deepfakes.

Moreover, I used Mann-Whitney U to analyze the relationships between the variables. I used the *p*-values to test whether there was a significant difference between two

variables. I wanted to test the relationship between gender and score, predicted score and score, confidence and score, and also whether the participants' knowledge of deepfakes will have an influence on the test score.

In N=92, for gender and score, the *z*-score is -1.78704. The *p*-value is 0.07346. The result is *not* significant at $p < 0.05$. For confidence and score, the *z*-score is 2.26863. The *p*-value is 0.0232. The result is significant at $p < .05$. For gender and predicted score, the *z*-score is -0.88936. The *p*-value is 0.37346. The result is *not* significant at $p < 0.05$. For knowing about deepfakes before and test score, the *z*-score is -0.99099. The *p*-value is 0.32218. The result is *not* significant at $p < 0.05$.

It seems the participants' confidence level to detect the deepfakes is significantly correlated with the ability to detect deepfakes ($p < 0.05$). Moreover, it stands to reason that the ability to detect deepfakes is related to the quality of deepfakes. The results seem to confirm that the technology of deepfakes is not mature at this point, at least in terms of the corpus studied, as it was easy for humans to detect deepfakes in this experiment. Among the 92 participants, the average test score is 7.61 /10. It means that among the 10 videos, among all of the 92 participants can guess more than 7 videos correctly. In addition, 84.8 % of the participants have a score more than 5.

RQ2

What are their reactions when deepfakes are revealed?

I tested emotional feedback before and after deepfakes were revealed. There were three things included - their concern towards the wide use of deepfakes, their feeling toward being deceived by deepfakes, and their trust in videos afterwards. Also the open-end questions about their feeling were collected after deepfakes were revealed. I used word clouds to measure emotional reactions.

From our experiment, we can see that for Figure 7, about whether or not deepfakes will deceive the participants, 48.9% of participants thought they were deceptive to some extent. 27.2% of participants took a neutral attitude toward it. Only 7.6% of participants thought it is not deceptive. For Figure 8, 54.3% of participants thought it will change their trust in video to some extent. 25% of participants took a neutral attitude. Only 2.2% of participants thought it will not change their trust in deepfakes.

For their concern towards increasing use of deepfakes, it seems most of participants showed concern about the increasing use of deepfakes regardless of whether it was before the test or in the middle of test or at the end of the test.

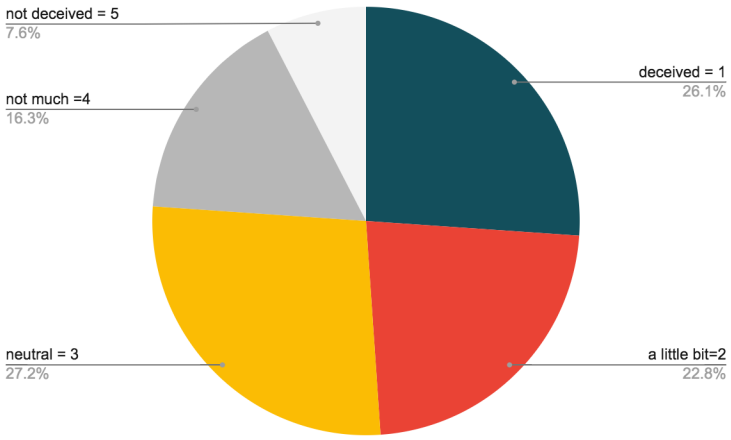

Figure 7. Chart of potential to be deceived by deepfakes

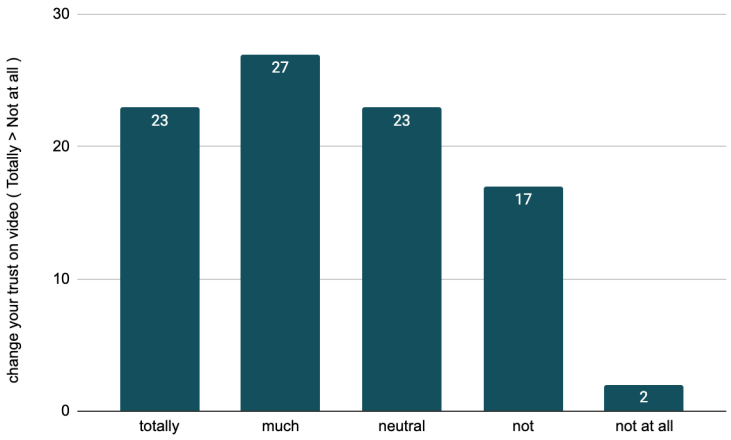

Figure 8. Change of trust in videos later

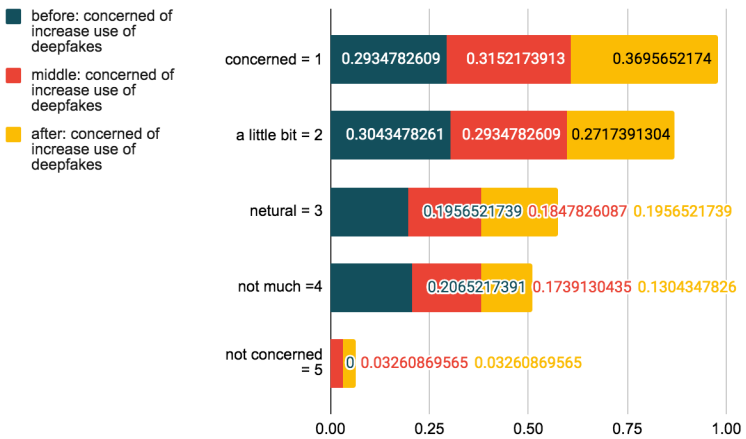

Figure 9. How concerned about the increasing use of deepfakes.

Moreover, I used two open-ended questions to test participants' emotional responses after they saw the deepfakes and after they got their test scores. I used word clouds to visualize the difference in their emotional responses before and after. From Figure 10 and Figure 11, we can see that they are often feeling good, interested, and happy before the test report and often feel good, happy, and confident after the test report.

Figure 10. Word cloud of their feeling before test report

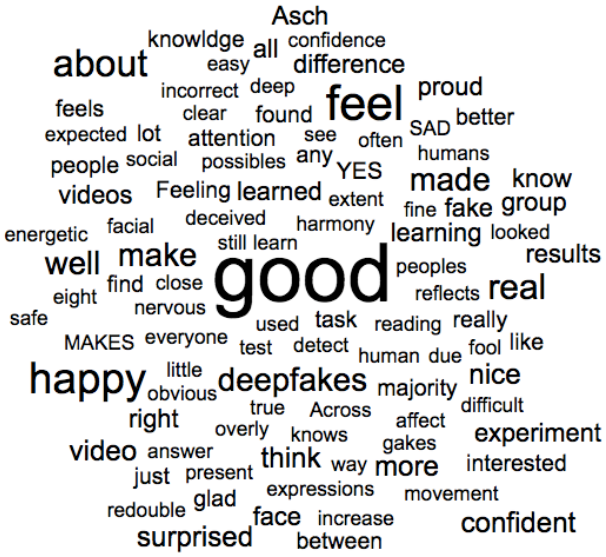

Figure 11. Word cloud of their feeling after test report

Before seeing the test score report, one participant said, “I feel good. It is an interesting study”. Another commented, “I just did a HIT before this watching 3 second clips

of a person talk and rating their emotion. But the whole clip is AI-made based on one image!! I imagine it is possible to create a deepfakes so good I as a common viewer could have no way to differentiate."

After test score report shows, they expressed, "it made me feel a little better and more confident." "It made me feel even more confident because I did well." Some showed some surprise towards their high score, "I was surprised I got eight correct."

RQ3

What do they see as the ethical implications of deepfakes, and how deepfakes could be used or abused?

There are several questions related to it and a PVQ value evaluation data will be collected to test whether or not there exists a relationship between their values and how they perceive the ethical implications of deepfakes. I used the 21-item Portrait Values Questionnaire (Schwartz, 2003) to measure 10 human values (Schwartz, 2007). I used the 10 categories of human values to test their relationships with test score. Also, some open-ended questions were devised to collect their feedback.

In PVQ questionnaire, Schwartz (Verkasalo, 2009) labeled these universal basic human values into 10 categories. They are benevolence, tradition, conformity, security, power, achievement, hedonism, stimulation, self-direction, and universalism. In these 10 categories, benevolence and universalism belong to Self-Transcendence Values, achievement and power belong to self-enhancement values, conformity, tradition and security belong to conservation value. Self-Direction and Stimulation belong to openness to change values. Self-Transcendence values are opposed to Self-Enhancement values. Conservation Values are opposed to Openness to Change Values. Hedonism value is located between openness to change and self-enhancement. In 21 questions, participants

were asked whether these statements are very much like me to not at all like me using a 6 point likert scale (see Table 1). All of these questions measure the 10 values listed in Table 2.

1: Thinking up new ideas and being creative is important to me.
2: It is important for me to be rich and have a lot of money.
3: I believe that every person in the world should be treated equally.
4: It is important for me to show my abilities. I want other people to admire what I do.
5: It is important for me to live in a safe and secure surrounding.
6: I love surprises and always want to try something new.
7: I believe that I should obey rules even when no one is around.
8: It is important for me to stay humble and modest.
9: I believe in listening to people who are different from me and try to understand them.
10: Having a good time is important to me. I like to 'spoil' myself at times.
11: I prefer to make my own decisions and do what feels right to me.
12: I like helping people around me.
13: Being successful is important to me.
14: It is important for me to ensure that the government is taking care of my safety concerns.
15: I want to take up new adventures and want to live an exciting life.
16: It is important for me to behave properly at all times and not do anything that people consider wrong.
17: It is important for me to earn respect from others.
18: Being loyal to my friends is a priority in my life.
19: I try to follow my traditional values and customs that my family and society have endowed on me.
20: I strongly believe that we should care about nature.
21: It is important for me to do things that give me pleasure.

Table 1: continued next page.

Table 1. 21 items of PVQ

BENEVOLENCE	12: I like helping people around me. 18: Being loyal to my friends is a priority in my life.
UNIVERSALISM	3: I believe that every person in the world should be treated equally. 9: I believe in listening to people who are different from me and try to understand them. 20: I strongly believe that we should care about nature.
SELF-DIRECTION	1: Thinking up new ideas and being creative is important to me. 11: I prefer to make my own decisions and do what feels right to me.
STIMULATION	6: I love surprises and always want to try something new. 15: I want to take up new adventures and want to live an exciting life.
HEDONISM	10: Having a good time is important to me. I like to 'spoil' myself at times. 21: It is important for me to do things that give me pleasure.
ACHIEVEMENT	4: It is important for me to show my abilities. I want other people to admire what I do. 13: Being successful is important to me.
POWER	2: It is important for me to be rich and have a lot of money.
SECURITY	5: It is important for me to live in a safe and secure surrounding. 14: It is important for me to ensure that the government is taking care of my safety concerns.

Table 2: continued next page.

CONFORMITY	7: I believe that I should obey rules even when no one is around. 8: It is important for me to stay humble and modest. 16: It is important for me to behave properly at all times and not do anything that people consider wrong.
TRADITION	19: I try to follow my traditional values and customs that my family and society have endowed on me. 17: It is important for me to earn respect from others.

Table 2. Ten Categories of PVQ

Table 3 lists the positive and negative perspectives on deepfakes expressed by participants. Overall, most participants show concern about misuse of deepfakes and most participants think the main positive aspect of deepfakes is for entertainment.

Good Perspective about Deepfakes	**Bad Perspective about Deepfakes**
I think this technology could be used for positive purposes, but the scenario is much more negative than positive.	I think it will serve exclusively to create false news and denigrate people. Viewers get misled by appearance, look

Table 3: continued next page.

We believe everything. Whether it's the news, a magazine, a book, gossip, it's all just words, spread around, by friends or "reliable" sources … There really is nothing good about deep fakes, it is cheating and I don't know why you would lie to people like this.	and voice of a faked person on TV. This person creates a false reality and viewers will be made even more unsure about what to believe on the Internet and on TV and what not. Certainly yes. There are many more harms. The main thing is the spread of false news among people with a harmful purpose. These videos can uncontrollably deceive and influence many people.

Table 3. Good perspective and bad perspective about deepfakes

Ethics Implications about deepfakes
People don't always consent to having their faces and voices used for deepfakes and the deepfakes can be used for nefarious reasons. Some sorts of restrictions will have to be imposed on when and how they can be used and offer punishment for those who don't

Table 4: continued next page.

use them within those boundaries.
Using an image of someone else without their permission, and using an image of someone to make it appear they said or did something they didn’t. I don't think deepfakes should be allowed.
I can't see a way to avoid this, perhaps an awareness campaign would be the best way to inform the person about this crime.
Deepfakes will serve to spread fake news and blackmail famous people and harm them.
I think people who have enemies or people who don't accept the end of a relationship can use this technique to get even.
Every service providers like youtube, facebook, twitter, etc., should monitor the video content and immediately remove the harmful content and suspend the TOS violated user accounts.

Table 4. Ethics implications about deepfakes

I also asked: “ What are the ethical implications of deepfakes, and what should be done to ensure that they are not abused?” I collected some answers from the participants in Table 4.

From these participants’ responses, I can see that before a new technology such as deepfakes are adopted into the real world, they should be well understood in order to ensure that it evolves on the right path. Laws should be used as a tool to regulate the misuse of new technology. Deepfakes, as a new technology, might be misused to mislead the masses and spread fake news for someone’s interest. Figure 12 shows the word cloud of answers about ethical implications.

Figure 12. Word cloud of ethical implications of deepfakes

The participants articulate significant concerns about deepfakes: “People don't always consent to having their faces and voices used for deepfakes and the deepfakes can be used for nefarious reasons. Some sorts of restrictions will have to be imposed on when and how they can be used and offer punishment for those who don't use them within those boundaries.” “Deepfakes will serve to spread fake news and blackmail famous people and harm them. I think people who have enemies or people who don't accept the end of a relationship can use this technique to get even.” “They can be used to spread misinformation or create panic due to creating fake news, and they are sort of inherently unethical in that regard. There should be mechanisms to detect deepfakes and delete those videos as soon as possible I think.” “If they're being used for nefarious reasons, then it would be unethical to use them. I don't know of anything you could do to stop people from using them unless you made it an actual crime, but even that wouldn't stop them.”

Moreover, I used the Mann-Whitney U to analyze the relationship between the data groups. I used *p*-values to test whether there was significance difference between each of the 10 categories of human values and the overall score.

The relationship between self-transcendence and score, the *z*-score is 1.0839. The *p*-value is 0.28014. The result is *not* significant at $p < 0.05$. For Self-enhancement and score, the *z*-score is 0.12175. The *p*-value is 0.90448. The result is *not* significant at $p < 0.05$. For conservation value and score, the *z*-score is 0.86796. The *p*-value is 0.3843. The result is *not* significant at $p < 0.05$. For openness to change, the *z*-score is -0.57002. The *p*-value is 0.56868. The result is *not* significant at $p < 0.05$. For hedonism, the *z*-score is -0.4127. The *p*-value is 0.6818. The result is *not* significant at $p < 0.05$. Therefore, human values were not predictive of test scores.

Discussion

Detection Ability

For RQ1, " How well do people detect deepfakes, and what factors affect their ability to detect deepfakes? " We can see that most of people can detect deepfakes and this is probably because the immaturity of deepfake technology. Most participants found it easy to recognize the deepfakes. It seems this experiment increased their confidence towards deepfake technology. It lowered their concern towards this new technology and assure their confidence in their ability to detect deep fakes. The confidence towards detecting ability had a relationship with their score. Due to the immaturity of deepfake technology, most of participants can perform well in detecting deepfakes.

People's Reactions

For RQ2, "What are their reactions when deepfakes are revealed?" Most of them will feel good towards deepfakes. Their concern towards deepfakes become lower when they get a higher score in the experiment and find they possess the ability to detect deepfakes easily. Most participants lost trust in any video information they saw later. Many of them think there should be some law to regulate the misuse of deepfakes. The test score made most of the participants feel good. Only a small proportion of participants felt surprised by deepfake technology.

Ethical Implications

For RQ3, "What do they see as the ethical implications of deepfakes, and how they could be used or abused? " From the open-ended question, we can see that more people are concerned about the bad influence of deepfakes, and they are worried about the fake news and the bad influence of its spread. They are worried about lack of regula-

tion towards deepfake technology and lack of punishment in relation to its misuse. On the contrary, for the most part, it seems most of people do not see any benefits of deepfakes. The main exception was the use of deepfakes in the entertainment industry.

Conclusion

Regulation Feasibility

At the end of 2019, there was a piece of news that Facebook will ban deepfakes ahead of the 2020 election. Facebook expresses their worries that deepfakes will mislead voters. Therefore, it is a trend that we should regulate the use of deepfakes before it is in widespread use (Makena, 2020). The reasons underlying it is that "a digitally manipulated video of House Speaker Nancy Pelosi falsely depicted her as slurring her words in a drunken manner during a public speech" (Drew, 2019, p. 1). This video was shared with millions of people and led to a heated controversy.

Influence on Humans

New technology will have an influence on humans, possibly for good and for bad. Even though most people can detect deepfakes, it does not mean they can handle this technology. From the experiment, participants increased their concern after they finish the experiment. It will influence their trust in videos. As a protection method for deepfakes (Güera, 2018), some systems using convolutional neural networks (CNN) have been created to detect whether the video is manipulated or not. Some technology (Hasan & Salah, 2019) is designed to use the smart contract in blockchain to trace the provenance of video. Through this way, we can trace the content to know whether it comes from a trusted source. These technologies may help us to avoid the bad influence of deepfakes on humans and reduce human worries towards misuse of deepfakes.

Trust in Information

Information is hard to trust in the modern world, whether audio, video, etc. New technologies can be appropriated to serve a wide range of interests. My findings under-

score the importance for people to cultivate a means of making informed trust judgments towards internet content such as news. We need to be able to tell apart the good things from the bad things. This skill is a critical competency in the digital era. Pan (2011) has shown that users can use negative comments to question the reliability of products. In addition, if the information is shared within close social relationships, it is easy for humans to build trust.

References

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, *1*, 316-334.

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

Drew, H. (2019). Faked Pelosi videos, slowed to make her appear drunk, spread across social media. Retrieved from https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/

Gambetta, D. (2000). Can we trust trust. Trust: Making and breaking cooperative relations, 13, 213-237.

Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research*, *5*(4), 269-306.

Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.

Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.

Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *Ieee Access*, *7*, 41596-41606.

Kramer, R. M. (2009). Rethinking trust. *Harvard business review*, *87*(6), 68-77.

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), 2056305120903408.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*.

Makena, K. (2020). *Facebook bans deepfake videos ahead of the 2020 election.* Retrieved from https://www.theverge.com/2020/1/7/21054504/facebook-instagram-deepfake-ban-videos-nancy-pelosi-congress

Ortony, A., Clore, G. L., & Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Ortony, A., Clore, G., & Collins, A. (1988). The cognitive structure of emotions. cam (bridge university press. *New York.*）

Pan, L. Y., & Chiou, J. S. (2011). How much can you trust online information? Cues for perceived trustworthiness of consumer-generated online information. *Journal of Interactive Marketing*, *25*(2), 67-74.

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

Sam, L. (2019). *The information: In Defense of Deep Fakes.* Retrieved from https://www.theinformation.com/articles/in-defense-of-deep-fakes

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. *Questionnaire package of the european social survey*, *259*(290), 261.

Schwartz, S. H. (2007). Value orientations: Measurement, antecedents and consequences across nations. *Measuring attitudes cross-nationally: Lessons from the European Social Survey*, 161-193

Verkasalo, M., Lönnqvist, J. E., Lipsanen, J., & Helkama, K. (2009). European norms and equations for a two dimensional presentation of values as measured with Schwartz's 21‐item portrait values questionnaire. *European Journal of Social Psychology*, *39*(5), 780-792.

Verma, N., Fleischmann, K. R., & Koltai, K. S. (2017). Human values and trust in scientific journals, the mainstream media and fake news. *Proceedings of the Association for Information Science and Technology*, *54*(1), 426-435.

Verma, N., Fleischmann, K. R., & Koltai, K. S. (2018). Demographic factors and trust in different news sources. *Proceedings of the Association for Information Science and Technology*, *55*(1), 524-533.

Verma, N., Fleischmann, K. R., & Koltai, K. S. (2019, March). Understanding online trust and information behavior using demographics and human values. In *International Conference on Information* (pp. 654-665). Springer, Cham.