CARNEGIE MELLON UNVERSITY

PROJECT REPORT

Deepfakes: How are we prepared for the post-truth era?

Authors:
Ruobing WANG (ruobingw),
Zihao ZHOU (zihaozho),
Ziyuan LIN (ziyuanli)

Mentors:
Prof. Norman SADEH,
Prof. Nicolas CHRISTIN,
Peter STORY

School of Computer Science

December 23, 2019

CARNEGIE MELLON UNVERSITY

Abstract

Deepfakes: How are we prepared for the post-truth era?

Deepfakes technology is by using computer vision and AI techniques to modify videos that makes someone say or do something he/she never said or did. While computer-generated virtual reality is nothing new, this emergent technology can produce fake videos that are increasingly resistant to detection. We intend to provide an in-depth assessment of the current deepfakes as well as the potential tools to cope with the malicious use of deepfakes. Therefore, we will try to answer how we are prepared for the deepfakes era in the four dimensions: Technological Hurdles, Credibility, Legal Solutions, Technological Solutions.

Chapter 1

Introduction

We are entering an era where people can freely create and tamper with information. In recent years, with the development of image modification technology, we can hardly believe what a picture tells us. But now, we are about to face a more powerful and terrifying technology that warns us we should not easily trust a video either.

Deepfakes are media that take a person in an existing image or video and replace them with someone else's likeness using artificial neural networks (Wikipedia contributors, 2019). This technology can be used to make a video of people saying something they would never say or doing something they would never do. It is already hard for a human to tell the unnaturalness of deepfake videos, and the technology is still improving, which will make it become more realistic and increasingly hard to detect.

Though deepfake technology might have some benefits, it can pose lots of harms.

For individuals, deepfakes can bring some serious trouble by producing someone's fake video or audio. Imagine one of your friends or families send you a video or a voice message asking for some money, or an email is sent to you attached with a video showing that the person you love is kidnapped, what would you do if you cannot contact that person because his/her phone was stolen? The terrifying part is, those things are not just imagination. Just a few months ago, a CEO was scammed out of $243,000 by a voice deepfake [1].

Meanwhile, making people do things they've never done or they will never do is a total violation of portraits. There are a bunch of deepfake videos of President Donald Trump can be found on YouTube [2], and most of deepfake videos are pornographic videos featuring women without their consent [3].

Considering a higher level of the problems that deepfake technology can bring, most of existing deepfake videos of President Donald Trump are making fun of him, people can easily tell those videos are fake, but what if someone makes a well-performed deepfake video of him and put that video on the Internet, how bad would that be? The impact of this video will increase exponentially over time. Even if that video can finally be proved fake, it will take a huge amount of time and effort to tell people what they saw was not real, and there is no guarantee that everyone who has seen the video will be told that video is fake. So a person who can produce well-performed deepfake videos can totally mislead people and disturb elections, this will cause great chaos in society.

[1] A Voice Deepfake Was Used To Scam A CEO Out Of $243,000. URL: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/1b3044c02241

[2] Donald Trump in Toddlers and Tiaras Deepfake. URL: https://www.youtube.com/watch?v=i9KrJFLYxTI

[3] Deepfake videos are a far, far bigger problem for women. URL: https://qz.com/1723476/deepfake-videos-feature-mostly-porn-according-to-new-study-from-deeptrace-labs/

Furthermore, many security systems use facial recognition, what if deepfake technology improved well-enough to fool those facial recognition algorithms in the future? If that happens, there will be a serious security problem.

More importantly, deepfake is not a complicated and fancy technology that only a little amount of professional people can use, there are a lot of easy-to-use tools and tutorial videos can be easily found. On China's biggest online shopping platform – Taobao, several groups of people are waiting to be paid to make deepfake videos, the only thing you need to do is providing some pictures or videos, and you can get a great deepfake video you want if you have several hundred dollars.

Although this technology has not been around for a very long time, the Deepfake-era is not far from us, we need to be alert before we are violated by this fast-growing technology, we need to come up with some countermeasures in advance to face the challenges that this technology may bring to our society.

To address the issue we are facing, our project mainly focus on four parts:

- **Technological implementation and hurdles of deepfakes**
  Making deepfake videos by ourselves, investigating different deepfake tools and estimating how much effort needed to be made for producing a deepfake video.

- **Credibility of deepfakes**
  Designing a survey to quantitatively measure the credibility of deepfakes and test human's ability of identify deepfakes.

- **Legal Solutions**
  Investigating existing statute and legislation attempts for deepfakes in the United States, trying to find out how are we prepared for deepfakes legally.

- **Technological solutions**
  Investigating different tools that are designed to detect deepfakes, measuring their abilities and performance to answer how are we prepared for deepfakes technically.

Chapter 2

Technological Implementation and Hurdles of Deepfakes

In order to better measure the difficulty of producing different levels of deepfake video, we have tried different kinds of deepfake technologies to generate our own deepfake videos. (The self-made deepfakes can also be used in our credibility survey and testing of existing deepfakes detectors) We also assess the difficulties of the different techniques, the time required for us to learn, and the quality of deepfake production of the different techniques.

2.1 Easy implementation of deepfake videos – DeepFaceLab

Firstly, we've investigated several tools to make deepfake videos. After estimating the effort we need to make, we chose the easiest one to start with.

DeepFaceLab is an easy-to-use tool that utilizes machine learning to replace faces in videos. This tool can be easily downloaded from their github web page.[1]Basically, this tool is a folder with a bunch of batch files. There are a lot of tutorial videos can be found on YouTube, some link of videos are also listed on the github web page. Figure 2.1 shows what it looks like. After putting the video resource in the folder, the only thing that needed to do is clicking those batch files step by step and always choose the default setting. For people who don't have a related computer science background, it will take more than an hour to learn and roughly understand how it works. If they want to run this tool on their own machine, they will need a GPU with at least 4GB memory, and it will take more than 3 hours to build the required environment.

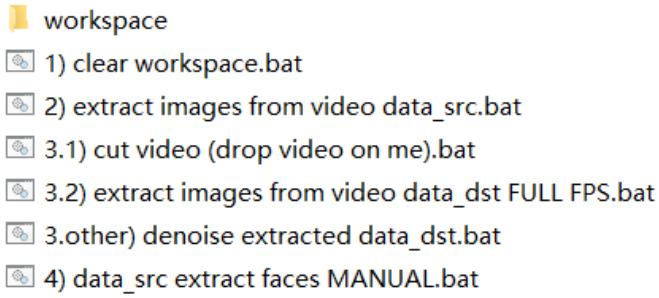
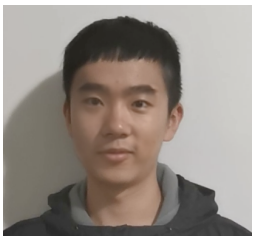


FIGURE 2.1: The interface of the DeepFaceLab tool

We've used two arbitrary videos that we could found to make our first deepfake video. We were trying to put Zihao's face in Ziyuan's video. After 8 hours of

[1]DeepFaceLab. URL: https://github.com/iperov/DeepFaceLab

(A) Arbitrary videos



(B) Specially-made videos

FIGURE 2.2: Final results with same tool but different video resources

training and 2 hours of parameter tuning, the final result we've got is shown as Figure 2.2a. Everybody can easily tell that this is not a regular video, this whole face doesn't even make sense, but that's the best we could do. Based on the quality of input-videos and the parameters we need to set for producing a deepfake video, we analyzed important features that have a great influence on the performance of the final deepfake video.

1. **Similarity between the shapes of two people's face**. Most deepfake tools are replacing face by only replacing facial futures, it would be really hard to change the shape of the face.

2. **Glasses**. For the same reason, if only one of the two people is wearing glasses, it's pretty hard to naturally remove the glasses or add a pair of glasses.

3. **Video resolution**. The video resolution determines how good the model will be trained and how good the final video will be. If the videos are really blurred or one of the videos is blurred, it would be really hard to produce a natural and convincing deepfake video.

4. **Size of faces**. If there is a great difference between the size of two faces in input-videos, there will be a great possibility that face in the produced deepfake video is unnatural. Just like the face in Figure 2.2a.

Based on the analysis above, we made two new videos as input. This time we carefully kept the light, the size of the face as similar as possible, and neither Ziyuan nor Zihao was wearing glasses. We also guaranteed the display resolution for both videos is 1920×1080. This time we were trying to put Ziyuan's face in Zihao's video because after we comparing the shape of their faces, we thought this way would be better.

We used the same tool, after 8 hours of training, with the default set of parameters, Figure 2.2b shows the final video we've got. There is nothing unnatural in this second deepfake video. We also tested how convincing this video was by asking Ziyuan to send this deepfake video to 4 of his friends and families without saying anything, it turned out none of them have raised any skepticism. Most of them just pointed out that Ziyuan had a new haircut.

2.2 Advanced implementation of deepfake videos – Faceswap

Faceswap is the leading free and Open Source multi-platform Deepfakes software. [2]

[2]Faceswap. URL: https://github.com/deepfakes/faceswapdeepfakesfaceswap

The top-level library API stack of this code is Keras which is a user-friendly library for programmers to implement and modify the structure of neural networks easily. The following lower-level backend that Keras interoperates with is Tensorflow developed by Google.

Here we briefly introduce the three processes of making deepfake videos with this Open Source code, i.e. extraction, training and converting.

The main idea of the extraction is to accumulate enough images of both sides of the people's face you want to swap or be swapped. For example, if you want to change a famous celebrity's face, the images can come from anywhere such as the photo taken at Cannes Film Festival, the screenshot of the film they were in and so on.

After we get a bunch of pictures of both sides, we start training the model. At a high level, training is teaching the neural network(NN) how to recreate a face. NN is made up of several models, and each model is largely made up of 2 parts:[3]

1. Encoder - this model takes a load of faces as an input and encode them into a latent state.

2. Decoder - this model takes latent states as an input and decode them into original faces.

We trained one model to encoded the first person's original face to a latent state and decoded this latent state to reconstruct the first person's face. Similarly, We trained another model to encoded the second person's original face to a latent state and decoded this latent state to reconstruct the second person's face. Finally, if we use the second model's decoder to decode the first model's output latent state, we will see the result that the first person's face is swapped to the second person's face.

Finally, we converted original video to deepfake video using the model we trained.

2.3 Technological Hurdles of Deepfakes

Due to the limitation of the easy implementation of deepfake video, we tried to make high-quality deepfake with following methods:

1. When we tried to produce real-life deepfake video(to modify the video we record by ourselves), we wore similar clothes, with no glasses, and made sure the angle we faced to the camera should not fluctuate too much.

2. For a video to be cut to generate images as training data set, we used the code to compute the similarity among the images and sorted them according to the similarity. Then, we checked the images cluster manually and deleted the images not suitable for training. This method greatly improved the performance of the final video.

3. We manually used a mask to mask the face according to the facial characteristic of the specific person and the background of the specific video.

We also measured the difficulties of the different techniques, the time required for us to learn, and the quality of deepfake production of the different techniques.

For different deepfake techniques, the length of learning time for us depends on the difficulty of the technique and whether existing a detailed tutorial and explanation of the theory. We learned how to implemented the following deepfake

[3]Faceswap. URL: https://forum.faceswap.dev

	Deepfacelab	Faceswap	Neural Textures	Face2Face	Faceswap + Fine-Tuning
Learning(h)	3	5	7	5	30+
Producing(h)	9	15	10+	10+	18+
GPU	GTX 1060	TESLA T4	TESLA T4	TESLA T4	TESLA T4
Performance	limited	good	good	good	perfect

FIGURE 2.3: The assessment of the effort we make to learn and implement the specific techniques and qualities of the different deepfake techniques

techniques including two we mentioned above: Deepfacelab, Faceswap, Neural Textures, Face2Face. we documented the training time and performance of each technology according to the video we made by ourselves and found on the internet with the same kind of GPU.

2.4 Conclusions

1. For people who have limited technical background, it is possible to spend less than 10 hours to learn an easy deepfake tool such as deepfacelab.
2. The better performance of the technique is, the more difficult for us to learn.
3. It is hard to make a convincing deepfake with limited image resources.
4. It is hard to make a well-crafted deepfake for a person who has no computer science background.
5. It is possible for sophisticated actors to make high-quality deepfake videos at a large scale given abundant good video resources.

Chapter 3

Credibility of Deepfakes

Deepfakes pose serious threats due to its inherent credibility which makes people believe in the algorithmically fabricated information. To further study this threat, it is necessary to quantitatively measure the credibility of deepfakes so that the seriousness of the deepfakes issue we are facing can be convincingly addressed. Thus, we designed and conducted a survey - Deepfakes Credibility Survey - to study this issue. This survey also aims to find what affects the credibility of deepfakes, which can provide insights for us to find ways to redress the threats of malicious deepfakes.

3.1 Design of the survey

Our goal of the survey is to measure the human's ability to identifying deepfakes as well as to find what could possibly enhance this ability. The details of the questionnaire can be found by scanning the QR code on the figure 3.1 or clinking the link. [1] The below will elaborate our design of the survey and why we design it in this way.

FIGURE 3.1: Scan the QR code to view our questionnaire

3.1.1 Subjects

Theoretically, our survey should be able to address different demographics to really measure the potential harm of deepfakes for our society. But as a course project, college students are the most available subjects we can find without financial compensation. Thus our subjects are 25 English-speaking college students currently studying in China or the US.

[1] Link to the survey. URL: https://qfreeaccountssjc1.az1.qualtrics.com/jfe/form/SV_87jCMWRP3IvivKl?Q_Language=EN

3.1.2 Videos

Although we have found large deepfakes datasets online and we also generated many well-crafted deepfakes of ourselves, the videos used for our survey is all focused on Barack Obama. We choose Obama is because he is the one that ubiquitously known by our subjects. Also, the video resources about him are abundant on the Internet, making it easier for creating well-crafted deepfakes. The reason why we keep using videos about a single person through our whole survey is to control variables. Because the survey has three parts. Thus we intend to keep the character in the videos consistent to make the comparison between parts of the survey meaningful.

The deepfakes are generated to make Obama say things that he did not say. However, we deliberately make the content of videos sounds normal since it is important that our subjects distinguish the deepfakes visually not by finding the deepfakes Obama sounds fake. The figure 3.2 shows the videos for the survey.

FIGURE 3.2: Videos for the different part of the survey

3.1.3 Procedures

Our survey intends to measure the human's ability to identified deepfakes in different situations so that we can analyse the results from different situations to find what could enhance the ability and redress the threats of deepfakes.

The first part of our survey intends to measure the daily life situation where people are not so vigilant that presume the videos he/she is watching could be a deepfake. The scenario is simulated by asking people to watch the first three of Obama's videos (including deepfakes) only once without telling them what the survey would be about. In this case, they are not warned the videos could be deepfakes thus it simulates the normal life situation. After they have finished watching the videos,

we asked them for each video if they can recall they have noticed the unnaturalness so that they raised skepticism for the videos.

For the second part, our subjects are aware that the survey is about deepfakes. If they are not, we explicitly introduce deepfakes and tell them the purpose of the survey. The survey will let them watch the other three of Obama's videos as many times as they want for them to catch if anything seems unnatural and raised their skepticism. In this case, they are warned they might watch a deepfake and allowed to try to distinguish it by watching many times. This scenario aims to measure the ability to identify deepfakes if people are aware of the existence of deepfakes and proactively seeking deepfakes. We also ask them to record how many times they have watched for each video on average.

The third part of our survey starts by telling our subject which video in the second part is a deepfake. We let our subjects learn the features of a deepfake by allowing them to watch and compare the videos in the second part as many times as they want. Our survey will record how the times they have watched for each video and how sure they are that they have caught the difference between a deepfake and a real video. After this, our subject is then asked to watch the other three of Obama's videos and try to distinguish the deepfakes. Therefore, these scenarios aim to measure to what extent people can learn to detect deepfake.

3.1.4 Summary

Goals: Measure the human's ability to identify deepfakes as well as to find what could possibly enhance this ability
Subjects: 25 English-speaking college students currently studying in the US or China
Video Materials: Nine videos of Barack Obama containing three of deepfakes
Procedures:

1. Without any disclosure of our study, ask subjects to watch three of Obama's videos (contain deepfakes) only once and see if they have raised skepticism for the deepfakes.

2. Tell the subjects the goal of our study, ask them to watch the other three of Obama's videos as many times as they want. Ask if they have raised skepticism for the deepfakes and record the time they have watched for each video on average.

3. Tell them which video in Step 2 is a deepfake, let them learn the features of a deepfake by watching the videos in Step 2 as many times as they want. Record how sure they feel about they have known how to distinguish deepfakes. And redo Step 2 with the new three of Obama's videos.

3.2 Results Analysis

From the survey data that we collected from the subjects, we can discover many interesting patterns and draw conclusions.

We first categorize the results from the subject into four categories.

- "Perfectly Classified" means the subject correctly identified deepfakes and real videos.

- "Identified" means they correctly identified deepfakes but not real videos, indicating they falsely marked a real video into a deepfake. This might be due to their over-skepticism about the videos.
- "Skeptically Identified" means they have chosen "Not sure" for the deepfakes, which indicates they somehow skeptically noticed something wrong in the deepfakes.
- "Not Identified" means they did not identify the deepfakes.

Using the above four categories, our results can be shown in figure 3.3 and figure 3.4. In part 1 of figure 3.3, there is 92.0% of subjects did not raise any skepticism for the deepfake. So it can be inferred that a well-crafted deepfake is extremely hard to be noticed if watched only once and not warned it could be a deepfake.

Also, there are still 52.0% of subjects that did not identify the deepfakes after being told that the goal of our study. So it can also be inferred that the effect of knowing the existence of deepfakes does matter. However, the human's ability to proactively identify deepfakes is still limited.

Comparing to the results of part 3, we can see that, even if the subjects were allowed to learn the features of a deepfake by watching the videos as many times as they want, the percentage of the people who cannot identify deepfakes remain the same. This indicates the insignificance of training a human observer to detect deepfakes.

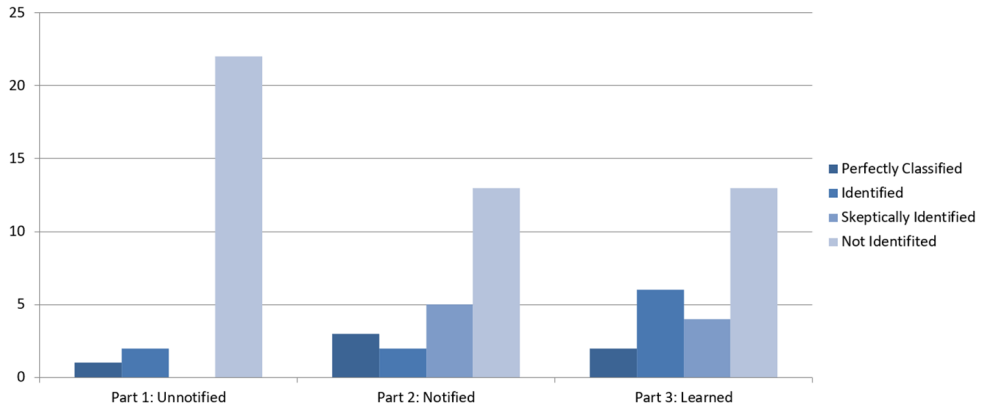


FIGURE 3.3: The result of identifying deepfakes on the three parts. The y-axis indicates the number of person on each category.

More interestingly, our survey asks the subjects in part 3 to record how sure they are about they have caught the difference between a deepfake and a real video. From figure 3.4, we can see that there is not a significant correlation between how the subject is sure about they have learned the difference and how the subject actually performed in identifying deepfakes. Thus it can be inferred that people might not be less vulnerable to deepfakes just because they are confident they have found the difference between a deepfake and a real video. This confidence might make them even more vulnerable to the harm of deepfakes because they are more likely to believe their own judgement of the authenticity of a video.

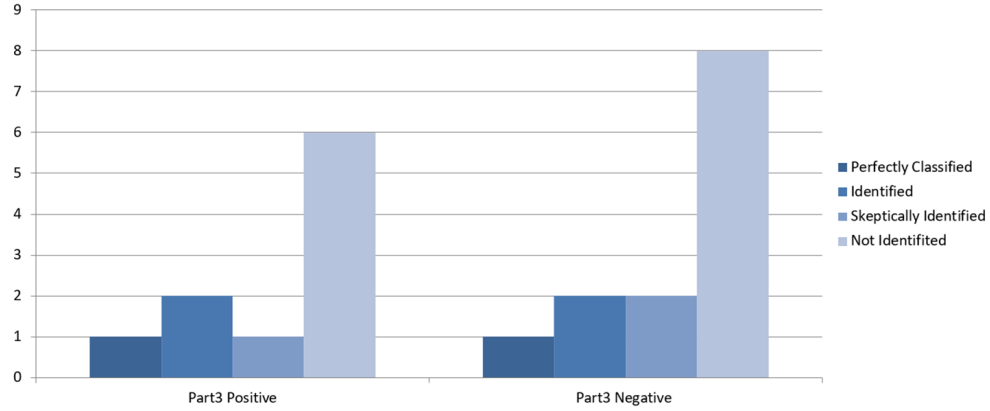


FIGURE 3.4: The result of identifying deepfakes after learned from examples. "Positive" (rate 4-10) means the subject is sure. "Negative" (rate 1-3) means the subject is not sure.

3.3 Conclusions

1. People are not likely to raise skepticism for a well-crafted deepfake if not warned.

2. Efforts to educate people to be careful about the existing of deepfakes are useful because people are more likely to raise skepticism to a deepfake.

3. The ability for human observers to proactively detect deepfakes is limited, so a technological solution for detecting deepfakes is necessary.

4. People might not be less vulnerable to deepfakes just because they are confident they have found the difference between a deepfake and a real video. This confidence might make them even more vulnerable to the harm of deepfakes because they are more likely to believe their own judgement of the authenticity of a video.

3.4 Limitations

Even we can draw many useful conclusions, our survey also has several limitations.

1. Our subjects do not address all demographic and the number of our subjects is limited.

2. The number of videos we used for the survey is limited to have a more generalized conclusion.

This limitation is mainly because we can not provide financial compensation to incentivize more subjects to participate in our survey and the time of taking the survey should be a restraint to less than 10 minutes for a voluntary survey.

However, our design of the survey is scalable. Essentially, our design of the survey remains the same with more subjects and more videos. Also, the hypothesis we tried to prove is the credibility of deepfakes. It is not quite necessary to generalize our conclusion to all deepfakes. We just have to measure what level of credibility deepfakes could possibly reach, since the sophisticated and malicious actors probably would use the most well-crafted deepfakes to reach its harmful purpose.

Chapter 4

Legal Solutions

When deepfakes came out with its striking ability to massively manufacture convincing fake videos and audios, it is natural to wonder if regulations are viable to cope with this challenge. While legislation may vary in different countries, this chapter will be mainly focused on the United States. By analyzing statute or legislation attempts, this chapter will try to answer the question: how are we prepared for the Deepfakes-era in the legal dimension?

4.1 Related Statute for Deepfakes

Deepfakes, as an emergent technology, has disrupted many traditional rules and beliefs. However, there are existing laws might be used against harmful deepfakes.

First, the violation of intellectual property can be used if it applies. Deepfakes requires original video resources to make fake videos. Some deepfakes might violate the copyright of the original videos by exploiting it for commercial purposes. In this case, the deepfakes can be taken down if the violation is found. However, this might not apply to the majority of the cases where the creators could claim that the deepfake is "fair use" of the intellectual property. Normally, "fair use" includes the following scenarios (the Copyright Act, 1976).

1. the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
2. the nature of the copyrighted work;
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. the effect of the use upon the potential market for or value of the copyrighted work.

Thus, the prospects for success by using the violation of intellectual property depends on the purpose of deepfakes and how much deepfakes has transformed from the original videos. Therefore, copyright infringement law has left much uncertain room for the defendant to argue its "fair use".

Second, there are also many tort laws could be used. The best fit is probably "false light" or defamation laws. There are 15 states have long-existed criminal defamation laws (Wagner and Fargo, 2015). False light or defamation can be used against deepfakes by making it a punishable offense to intentionally spread false information about a person. Also, plaintiffs could use the tort of Intentional Infliction of Emotional Distress (IIED), based on the proof of "extreme and outrageous conduct". [1]

[1] Benjamin Zipursky: Snyder v. Phelps, Outrageousness, and the Open Texture of Tort Law. URL: http://via.library.depaul.edu/cgi/viewcontent.cgi?article=1136/context=law-review

Moreover, the tort of "right of publicity" is also available if the deepfakes exploited one's right of publicity for commercial gain.

The currently available laws for suing deepfakes are substantial to some extent. The creators of malicious deepfakes can be held accountable for the claims, such as the violation of intellectual property, false light or defamation, intentional infliction of emotional distress, and right of publicity. However, those laws all come with limited coverage, which makes them not sufficient and robust enough to redress a broader range of harms deepfakes are posing. In this respect, we need new legislation effort to find a way to cope with this issue efficiently.

4.2 Legislation for Deepfakes

The legal effort against disinformation is nothing new in the US. However, deepfakes are the emerging weapon that can efficiently alter truth and garner attention. Especially on the Internet, the study shows that social networks incentivize virality, not veracity (Vosoughi, Roy, and Aral, 2018). The fact that virality can overwhelm truth has elicited swift responses from lawmakers in the fear of deepfakes disrupting the 2020 presidential election.

The first response is S.3805 - the Malicious Deep Fake Prohibition Act (Ben Sasse, 2018), which proposed to prohibit intentionally create or distribute a deepfake that "facilitate criminal or tortious conduct under Federal, State, local, or Tribal law." The definition of deepfakes in this bill is "an audiovisual record" that "would falsely appear to a reasonable observer to be an authentic record." However, this definition has a pitfall, that if a disclaimer is added to a deepfake, the subject of the deepfake could still be harassed or embarrassed, even the disclaimer makes it appear fake to the "reasonable observer."

Unlike the first legislation attempt which implicitly requires a disclaimer for a deepfake, H.R.3230 - the DEEPFAKES Accountability Act, proposed in June 2019, explicitly makes it a crime to create deepfake without adding "irremovable digital watermarks, as well as textual descriptions" (Yvette D. Clarke, 2018). Obviously, this act shares the same loophole as the S.3805 has.

While legislation at the federal level is not yet successful until now, legislation on the state level regarding deepfakes has already become effective. Assembly Bill No.602 has been approved by California governor Gavin Newsom, creating a private right of action against knowingly creating or distributing an "altered depiction" (Deepfakes) made without the person's consent (Leyva, 2019). Also in California, Assembly Bill No.730 criminally prohibits a person from producing or distributing with the intent to maliciously target a political candidate "in an election that is occurring within 60 days" (Grayson, 2019).

4.3 The Challenges of Regulations

In the first section of this chapter, it is concluded that, while the current laws can be used for suing the creators of deepfakes to some extent, it has limited coverage. Facing this new technology, it requires us to legislate new laws to cope with it. Although lawmakers' haste to regulate deepfakes has already been written on the two federal bills, it is indeed challenging for legislators to define what type of deepfakes is harmful and should be regulated, to find a way to efficiently reduce the amount of harmful deepfakes, to reduced malicious deepfakes online without undermining

vitality of the Internet. However, there are inherent obstacles to legal regulations to efficiently address the problem of deepfakes.

First, it is challenging for legislators to decide the **exact definition** of harmful deepfakes because failing to do so could undermine the First Amendment. For example, there is a deepfake about Mark Zuckerberg bragging about abusing data privacy.[2] This satirical deepfake brought public attention to the data privacy violation in Facebook, sparking social-valuable discussions. But all the federal level legislation (S.3805 and H.R.3230) failed to address the difference between malicious deepfakes and satirical deepfakes. Thus, it is natural to worry that by regulating all use of deepfakes as a whole would undermine the First Amendment and discourage social-valuable expression like the Mark Zuckerberg deepfake.

Notably, even if legislators find a way to define malicious deepfakes, the algorithmic way for the platform to distinguish satirical deepfakes from harmful deepfakes is almost impossible to be implemented in the foreseeable future. However, if malicious deepfakes could not be taken down by platform quickly in an automatic way, the potential damage could still be imposed on society. In this respect, not only it is challenging to come up with an accurate definition, but also it is almost impossible for the online platforms to consistently enforce the definition, making it extremely difficult for the platforms to cooperate with the regulations.

Second, if the lawmakers successfully addressed the real harmful deepfakes to regulate, the main obstacle is the **provenance** of malicious deepfakes since it is inherently hard to trace back to its creators. The bill S.3805 tried to regulate deepfakes by making a watermark disclaimer mandatory. This is understandable because for any regulations could not take effect if the provenance of deepfakes is insufficient. However, enforcing a watermark to deepfakes cannot efficiently solve this issue. Watermarks are easy to remove even if the creator tried to make it irremovable. With video editing tools, watermarks can be a mask or cropped by a few clicks of the mouse. As a result, the creator of a deepfake has no way to make sure its deepfake remained abiding by the bills (S.3805 or H.R.3230) after it was posted online. More importantly, those who are willing to add watermarks are probably the creators that the public need not to concern. But it is hard to imagine the bill could make those who "maliciously" intent to use deepfake for criminal activity put a watermark. If they were willing to add a disclaimer, they wouldn't use deepfakes in the first place. The consequence could be even worse. The fact that this watermark bill forces many innocuous deepfakes to carry a watermark, could even make people less vigilant to the video without a deepfake watermark, making harmful deepfakes even more dangerous. Especially, the sophisticated but malicious actors could use tools like Tor to make themselves anonymous. In this case, they might be impossible to find by any means.

Third, if the above two problems have solved, another obstacle is the **global nature** of online platform (where the deepfakes could circulate) and the **limited reach** of domestic regulations (where laws could regulate).

Fourth, even if the creators can be found and reside in the US, the **cost of suing deepfakes** could be both financially and mentally heavy, making the victims reluctant to pursue a lawsuit. Deepfakes can sabotage one's reputation and be extremely embarrassing. Thus pursuing a lawsuit brings publicity which might exacerbate the victim's harm.

[2]Deepfake about Mark Zuckerberg. URL: https://www.washingtonpost.com/nation/2019/06/12/mark-zuckerberg-deepfake-facebook-instagram-nancy-pelosi/

Last but not least, while the above obstacles are all focused on seeking accountability for the creators of deepfakes, more significant obstacles shield the content platforms, which enables the circulation of deepfakes, from being responsible for the propagation of malicious deepfakes. The **Section 230 of the Communications Decency Act (CDA)** provides immunity for the online platforms to host harmful content that created by its users, unless the content violates federal criminal law, intellectual property law, and the Electronic Communications Privacy Act. Basically, the internet is mostly left to govern itself, outside of the three exemptions.

4.4 Conclusions

It is perfectly normal that an emergent technology disrupts the current rules and beliefs, creating a legal gap that requires years to fill. In essence, the real problem lawmakers need to efficiently address is **the weaponization of deepfakes**. The key point is to regulate the malicious use of deepfakes as a weapon that poses social harm without undermining the freedom of speech and the vitality of the Internet.

So the conclusion in the legal dimension is, due to the inherent obstacles for regulating deepfakes discussed above, the current laws and attempted legislation could not solve this problem comprehensively and accurately. In this respect, we are not currently well-prepared for the legal threats deepfakes are posing to our society.

Chapter 5

Technological Solutions

If laws alone cannot ameliorate all the harms of malicious deepfakes, might technology? Technology advance in Artificial Intelligence is the crucial factor that deepfakes thrived. It is natural to wonder if technology can solve the problems it has created itself.

5.1 Existing Efforts and Methods

Given the serious threats of deepfakes, the Internet giants company is under pressure to find a way to tackle the deepfakes. Google has released a large deepfakes data set which provides a testing environment for the people to invent efficient methods to detect deepfakes.[1] Also, Facebook and Microsoft have teamed up to launch a deepfake detection challenge.[2] Not only private sectors tried to tackle deepfakes technically, but the US government also steered funds towards this issue by launching the DARPA's "Media Forensics" project.[3]

Following the trend, there are many academic papers that came out and try to detect deepfakes. Xception is a forgery detection method that detects deepfakes by use additional domain-specific information (Rössler et al., 2019). MesoNet is proposed to use two deep neural networks to detect deepfakes with the capability to mitigate the affects from the compression of a video (Afchar et al., 2018). There is also a deep learning method to detect deepfakes by adding a novel convolutional layer (Bayar and Stamm, 2016). Similarly, a new deep learning method has been purposed by using a custom pooling layer as an ensemble method to different deepfakes detection algorithms (Nicolas Rahmouni and Echizen, 2018). Also, a detection method involving residual-based descriptors can be used as simple constrained convolutional neural networks (Cozzolino, Poggi, and Verdoliva, 2017). In addition, there is also a novel deepfakes detection algorithm that used a novel method of building steganography detectors instead of using neural networks (Fridrich and Kodovsky, 2012).

The Google Deepfake Data Set is widely used for today's deepfakes detection benchmark. Thus we tried to evaluate the open-source algorithms mentioned above along with the non-open-source benchmark evaluated by others. The benchmark is showed in figure 5.1.

[1] Google Deepfake Data Set. URL: https://github.com/ondyari/FaceForensics/

[2] The Facebook and Microsoft Deepfake Detection Challenge. URL:https://ai.facebook.com/blog/deepfake-detection-challenge

[3] DAPRA's "Media Forensics (MediFor)" project. URL: https://www.darpa.mil/program/media-forensics

	Deepfakes	Faceswap	Face2Face	Neural Textures
Xception	0.952	0.901	0.851	0.803
Xception Full Image	0.873	0.869	0.903	0.807
MesoNet	0.742	0.705	0.746	0.733
Bayar and Stamm	0.845	0.825	0.737	0.707
Rahmoni	0.855	0.563	0.642	0.607
Recasting	0.855	0.738	0.679	0.780
Steganalysis	0.736	0.689	0.737	0.633

FIGURE 5.1: The accuracy of detecting different deepfakes by different detection algorithms (Rössler et al., 2019), (Afchar et al., 2018), (Bayar and Stamm, 2016), (Nicolas Rahmouni and Echizen, 2018), (Cozzolino, Poggi, and Verdoliva, 2017), (Fridrich and Kodovsky, 2012).

5.2 The Challenges of Technological Solutions

The above section introduced the existing technological solutions to detect and debunk deepfakes. Though the works may seem abundant, there are still many obstacles for us to redress the harm of deepfakes efficiently.

First, the current detection algorithms are not able to generalize well on the unforeseen deepfakes, making the current solutions neither scalable nor reliable. The models in figure 5.2 are all fine-tuned on the Google Deepfakes Data Set. While it might be able to reach comparatively decent accuracy, we used the exact model to detect many self-made deepfakes of our own and found the model not reliable. We are able to find many cases where the model predicted a deepfake as a real video or predicted a real video as a deepfakes. The figure 5.2 shows a incorrectly classified example.

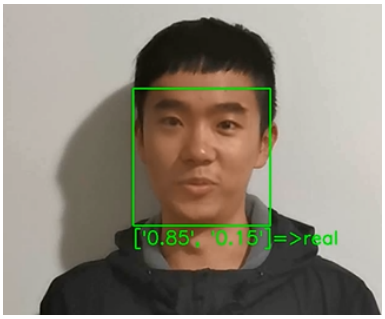


FIGURE 5.2: Our self-made deepfakes detected as a real video.

Increasing the accuracy of the detector and the ability of generalization is a classic problem in machine learning. Although our current model might not be the perfect solution, we can be optimistic about we could reach a much more accurate method in the foreseeable future, just for the four deepfakes algorithms listed above.

However, the real issue of developing a deepfakes detector is to cope with the "arms race" between a detector and a new deepfakes algorithms. We can already notice that the existing tools perform better on the traditional Deepfakes algorithm (see the Deepfakes column on the figure 5.1), but worse on the newer Neural Textures (see the Neural Textures column on the figure 5.1). Whenever there is a new

detector of deepfakes, it gives a new discriminator for the attacker to design and fine-tune his deepfakes specifically against the weakness of the detector.

Moreover, even the online platform managed to detect deepfakes, there is no algorithmic way to decide whether the platform should take down the deepfakes. As we mentioned in Chapter 4, an automatic way to decide if a deepfake is satirical or harmful is beyond the reach of current AI capability. While it helps to redress the harms of deepfakes if detectors can automatically mark deepfake, it still requires human effort (i.e. human content moderator) to respond to harmful deepfakes.

5.3 Conclusions

1. The current technological solutions are neither scaleble nor reliable enough to mitigate the harms of malicious deepfakes.
2. The development of deepfakes detectors suffers from the "arms race" with the same rapid revolution of new deepfakes technology.
3. Even the online platform managed to detect deepfakes, there is no algorithmic way to decide whether the platform should take down the deepfakes.

Chapter 6

Conclusions

According to the four parts of the investigation we've made, we can draw the conclusion as follows.

1. Technological hurdles are easy to overcome by sophisticated actors which can produce high-quality deepfakes on a large scale.
2. It is hard for people to raise skepticism for good deepfakes with no warning ahead.
3. There are inherent obstacles for regulating deepfakes either by the current statutes or new legislation, making the legal gap challenging to fill.
4. The current technological solutions are neither scaleble nor reliable enough to mitigate the harms of malicious deepfakes.

In the final analysis, we are definitely not currently well-prepared for the harms that deepfakes poses to us. Calibrated solutions might require both legislators and technologists to work corporately. However, it is perfectly normal for an emergent technology to disrupt our traditional rules and beliefs, leaving gaps for years to fill. There are indeed many obstacles for us to redress the harms of deepfakes efficiently. But by addressing the seriousness sufficiently and bringing public attention, there would be more public pressure and market drive for more people to come up with innovative solutions to this disruptive technology. In this case, we will be eventually well-prepared.

Bibliography

Afchar, Darius et al. (2018). "MesoNet: a Compact Facial Video Forgery Detection Network". In: *arXiv e-prints*, arXiv:1809.00888, arXiv:1809.00888. arXiv: 1809.00888 [cs.CV].

Bayar, Belhassen and Matthew C. Stamm (2016). "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer". In: *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*. IH&MMSec '16. Vigo, Galicia, Spain: ACM, pp. 5–10. ISBN: 978-1-4503-4290-2. DOI: 10.1145/2909827.2930786. URL: http://doi.acm.org/10.1145/2909827.2930786.

Ben Sasse (2018). *S.3805 - Malicious Deep Fake Prohibition Act of 2018*. https://www.congress.gov/bill/115th-congress/senate-bill/3805/text?format=txt.

Cozzolino, Davide, Giovanni Poggi, and Luisa Verdoliva (2017). "Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection". In: *arXiv e-prints*, arXiv:1703.04615, arXiv:1703.04615. arXiv: 1703.04615 [cs.CV].

Fridrich, Jessica and Jan Kodovsky (2012). "Rich Models for Steganalysis of Digital Images". In: *IEEE Workshop on Information Forensics and Security,*

Grayson (2019). *Assembly Bill No. 730*. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730.

Leyva (2019). *Assembly Bill No. 602*. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602.

Nicolas Rahmouni Vincent Nozick, Junichi Yamagishi and Isao Echizen (2018). "Distinguishing computer graphics from natural images using convolution neural networks". In: *IEEE Workshop on Information Forensics and Security,*

Rössler, Andreas et al. (2019). "FaceForensics++: Learning to Detect Manipulated Facial Images". In: *arXiv e-prints*, arXiv:1901.08971, arXiv:1901.08971. arXiv: 1901.08971 [cs.CV].

the Copyright Act (1976). *Section 107 of the Copyright Act*. https://www.copyright.gov/title17/92chap1.html#107.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). "The spread of true and false news online". In: *Science* 359.6380, pp. 1146–1151.

Wagner, A. Jay and Anthony L. Fargo (2015). *Criminal Libel in the Land of the First Amendment*. http://legaldb.freemedia.at/special-report-criminal-libel-in-the-united-states/.

Wikipedia contributors (2019). *Deepfake — Wikipedia, The Free Encyclopedia*. [Online; accessed 7-December-2019]. URL: https://en.wikipedia.org/w/index.php?title=Deepfake&oldid=929567888.

Yvette D. Clarke (2018). *H.R.3230 - Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019*. https://www.congress.gov/member/yvette-clarke/C001067.