# NYS Senate Standing Committee On
# Consumer Protection And Internet & Technology
# Protecting Consumer Data and Privacy on Online Platforms, 11/22/2019

Siwei Lyu, Ph.D.
University at Albany, State University of New York

**What are DeepFakes?**
DeepFakes are realistic human faces, voices, and activities that are created with the help of the advanced AI, and in particular, deep learning technologies. They are the recent twists to the disconcerting problem of online disinformation. The term *DeepFake* first emerged in late 2017 as the name of a Reddit account that began posting synthetic pornographic videos generated using an AI-based face-swapping algorithm. The term has subsequently become synonymous with AI-generated impersonating images, audios, and videos.

**Why we should care about DeepFakes?**
Photos and videos have been doctored since their nascence. But there are three reasons why the current concerns over DeepFakes and other AI-generated fake media are justified.

- **First, DeepFakes can be made more easily, quickly, and with better quality** — thanks to the rapid advancement of computer hardware and software technology, in particular those related to AI. For instance, to create a to create a realistic face-swapping DeepFake video, one only need to have a computer equipped with a special computing hardware known as graphical computing unit (GPU) and an Internet connection, which can be purchased for an affordable price on Amazon.[1] There are also websites that provide generation of fake media as a paid service, e.g., Deepfakes Web β (https://deepfakesweb.com).
- **Second, the capability to make DeepFake has been democratized through numerous software tools that can be downloaded freely from online code sharing platforms.**[2] These tools have made the process fully automatic and streamlined, so anyone with basic knowledge on this subject and required hardware/software can potentially make DeepFakes.
- **Third, with the abundance of online media we share, anyone is a potential target of a DeepFake attack.** A fake video showing a politician engaged in an inappropriate activity may be enough to sway an election if released close to voting day. A fake video of a falsified recording of a high-level executive commenting on his/her company's financial situation could potentially send the stock market awry. A fake video made by falsely implanting a woman's face in a pornographic video and shared on social-media platforms could tremendously traumatize the victim. A recent incident of a fraudulent money transfer initiated by a synthesized voice of a high-level executive to an employee has led to real financial consequences to the company.

[1] An example of computer configuration for this purpose includes an HP-Z800 workstation (~$1,000) equipped with an Nvidia 2080Ti GPU (~$1,200) and other necessary peripherals. Cost effective and large-scale production can also be conducted using cloud platforms such as Amazon AWS or Google Cloud Platform.

[2] e.g., FakeAPP (used to be on Reddit but now defunct), DeepfaceLab (https://github.com/iperov/DeepFaceLab), faceswap-GAN (https://github.com/shaoanlu/faceswap-GA), faceswap (https://github.com/deepfakes/faceswa), DeepFakeLab (https://github.com/iperov/DeepFaceLab), DFaker (https://github.com/dfaker/df), and more recently ZAO (https://apkproz.com/app/zao).

There is also a growing number of online fake pornographic videos generated using AI-algorithms[3]. The stakes are too high to ignore.

**How are deepfakes made?**

Deepfakes are created with a type of AI technology commonly known as deep neural networks. For instance, a deep neural network model learns to synthesize realistic faces and voices through *training*, which involves exposing the model to a large number of face images and voices of different subjects. Once the model is properly trained, it is ready to be used to generate DeepFakes. Current computer hardware and AI technology has made it much easier to create deepfakes. The training videos for the targets can be easily downloaded from social-media platforms such as YouTube, Instagram, and Facebook in large volume and high quality. Convenient software tools have made the whole process automated barring the choice of a few parameters. As a result, a few good-quality, minute-long videos, a commodity computer with a GPU, and several hours of training are sufficient to generate fake media with good visual quality. The fake media can then be distributed onto various social platforms and spread rapidly.

**How does technology combat DeepFakes?**

It is thus important to develop effective technologies to identify, contain, and obstruct deepfakes before they can inflict damage. There are various technical developments to combat fake media.

**Detection**: Effective DeepFake detection methods look for such traces to differentiate deepfakes from real videos. For instance, synthesized faces are warped and processed to fit the target's head orientation, such operations leave traces that can be exploited to detect deepfakes. Another type of detection techniques involves examining physiological inconsistencies such as the lack of realistic eye blinking and heart beating. A third approach is to "use AI to fight AI", using another deep neural networks to detect deepfakes. State-of-the-art detection methods have shown some promising accuracy on benchmark datasets but their actual performance on real life deepfakes have yet to be tested.[4] Nevertheless, the competition between making and catching fake media is an ongoing battle, just as digital media forensics can use nuances in the synthesis algorithms to develop better detection algorithms, forgery making algorithms can also benefit from learning how the detection works and continue to improve. As such, no matter how sophisticated detection algorithms are, there will always fake media that elude the detection. The goal of developing detection methods is to raise the bar on technical skills, time, resources, and efforts to make undetectable fake media.

**Protection**: In addition to detection methods, we also need methods to protect personal privacy and slow down training of AI models. My lab is experimenting with one such technology that can prevent the re-use of online images and videos as training data for generating fake videos. This involves inserting imperceptible "adversarial noise" into images and videos before they are uploaded to online social-media platforms. The adversarial noise correspond to subtle perturbations that human eyes cannot see nonetheless can disrupt a face detection algorithm and make it difficult to automate the training process. A dedicated adversary could overcome

[3] According to the report in September 2019 by the Dutch Company DeepTrace, the total number of online DeepFake videos is close to 15,000 with more than 130 million views, 96% of these videos are pornographic in nature and the subjects involved are 100% women. Source: https://deeptracelabs.com/mapping-the-deepfake-landscape/

[4] One notable effort towards this goal is the upcoming *Deepfake Detection Challenge* (https://deepfakedetectionchallenge.ai) sponsored by Facebook, Microsoft and Partnership on AI, to advance the state-of-the-art deepfake detection capacities.

adversarial noise by painstakingly selecting the target's face in every frame of a training video, but that requires 1,500 hand-marked selections for each 60 second training video.[5]

**Authentication and Verification**: There are also technologies that can tell us what is real instead of what is fake. For instance, digital watermarks can be inserted into authentic media and later verified to prove no manipulation has taken place. Control-capture technologies can authenticate content by extracting, at the time of capture, a unique digital signature from the medium, and then placing the encrypted signature on a secure central server or a distributed immutable ledger system such as the blockchain. The signature can later be retrieved to compare with versions of the same medium to determine if it has been changed since the time of capture.

### How to combat DeepFakes more effectively?

However, fighting DeepFakes is not only a technical problem. We need a comprehensive and robust solution to this problem beyond developing counter technologies. Below are some actionable items involving a broader community.

- First, without the collaboration of the platform providers from the technology sector (including, Facebook, Google/YouTube, Instagram, and Twitter), the protection brought forth by the development of counter technologies will not benefit the public. The major technology companies must more aggressively and proactively deploy technologies to form the first line of defense against AI-generated fake media.
- Second, we should also break the ecosystem and financial incentives of making malicious AI-generated fake media. Currently, making fake media is greatly facilitated by the lack of regulations on open-source free software tools making them, the abundance of personal data we share online, and the high financial incentives in providing such service.
- Third, we should increase investment in research programs that seek to build counter technology to fake media, and provide more incentive to startups that bring such counter technology to the broader public.
- Last but not least, due vigilance from the public is the best defense to the problem caused by DeepFakes and other AI-generated fake media. To this end, we need to educate the public on how to consume trusted information, on how to be better digital citizens, and on how not to fall victim to scams, fraud, and disinformation.

### Conclusions

It is not an exaggeration to say that we are on the cusp of deepfakes being cheap, easy to produce, indistinguishable from real videos, and ready to cause real damages. We therefore need a comprehensive and robust solution to this problem. The situation calls for a joint force from the government agencies, platform companies, media outlets, technical and research community, as well as the ordinary online users. Finally, the situation surrounding DeepFakes may not turn out to be as apocalyptic. But it is better safe than sorry.

[5] This is calculated based on a target video quality of 25 frames per second, which is the lowest frame rate for YouTube videos.