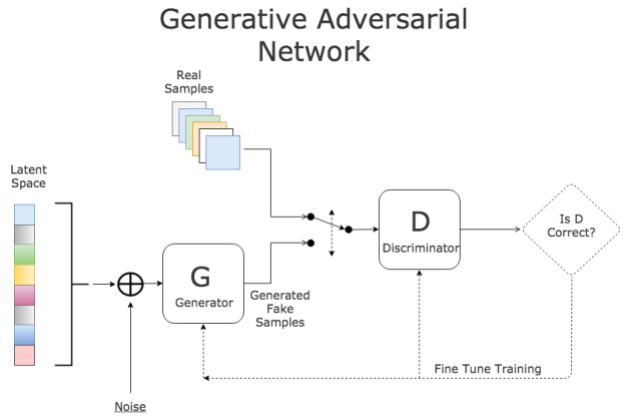
| | |
|---|---|
| **FORUM:** | United Nations Commission on Science and Technology for Development |
| **ISSUE:** | Measure to Address the Issues Rising from the Recent Surge in Deep Fake Video and Audio Technology |
| **STUDENT OFFICER:** | Gyuyeon (Nicole) Choi |
| **POSITION:** | President of United Nations Commission on Science and Technology for Development |

---

## Introduction

The term “deep fake” is a combination of words “deep learning” and “fake” and it refers to any visual or aural content produced by artificial intelligence (AI) that impersonates a targeted individual. This deep fake technology employs a subfield of AI called machine learning (ML): generative adversarial networks (GANs), to be precise. In the operation of GANs, two machine learning systems –neural networks– are trained against one another as competitors. While the generator aims to produce counterfeit data that reproduces the quality and features of the original data provided, the discriminator aims to distinguish the replica from the original. The generator makes alterations accordingly at each iteration until the discriminator is unable to tell apart between the original and counterfeit data – it may take up to thousands or even millions iteration until this is achieved.




*A visual representation of the generative adversarial network (GAN)*

Today, majority of the deep fake videos and audio recordings can often be detected without the need for sophisticated detection tools. However, as AI technology is improving at a remarkable rate as never before, the manipulated audiovisual contents are becoming increasingly convincing, accessible, and easy to process. Therefore, deep fakes pose a serious threat to the credibility of the sources online and the security of information of an individual or a group.

# Background

*Deep fake video of Obama warning the audience of the emerging dangers presented by the deep fake technology*

Though the mechanism behind the technology has been developed since the 1990s, the term "deep fake" was coined in 2017 by a Reddit user of the same name who created and shared AI generated sexually explicit media of celebrities. Now, "deep fake" has broader implications and is synonymous with numerous problems originating from the internet such as fake news, disinformation, and media manipulation. This technology only started to gain attention from the public back in 2018 when Buzzfeed released a video called *You Won't Believe What Obama Says in This Video!* As the title of the video suggests, it featured the former U.S President Obama warning the viewers that "We're entering an era in which our enemies can make it look like anyone is saying anything at any point in time" and he was presented to be saying things that he would have never said, at least publicly. The video soon revealed that it was, in fact, a manipulated video created by deep fake technology with the voice impersonated by an actor Jordan Peele. After it was uploaded, it stupefied the users on internet communities because of how convincing the video initially seemed without knowing that it was fabricated. In 2021, the video exceeds 8.6 million views on YouTube which goes on to show the impact it had upon raising awareness on the dangers of AI technologies.

# Problems Raised

## *Creation and Distribution of Sexually Explicit Media*

In a survey conducted in September 2019 by DeepTrace, it was revealed that 96% of the deep fake videos that were found on the internet were pornographic. Though it originally started off as pasting faces of celebrities into porn videos of people's choice, it is now expanding to victimize a wider group of individuals, especially women. People are stealing women's faces online to make sexually explicit media that the subjects never consented to be in.

Now, editing and producing convincing videos with someone else's facial features requires an extensive training of software on hundreds and thousands of images of the same person with different expressions, angles, and lighting until the computer thoroughly learns the face of an individual at different situations. That is why celebrities were the first to fall victim to the misuse of deep fake

videos because a plethora of images has long been available of public figures online. In the recent years, the collection of such data has not been a difficult task for anyone with a strong social media presence, which resulted in random targeting of ordinary users to unknowingly fall victim as well.

Though large streaming sites do have policies in place banning sexually explicit deep fake content, these videos seem to proliferate, nonetheless.

## *Financial Cyber Fraud*

*A depiction of cybercrime using forged audio to defraud money*

Ever since the emergence of deep fake video and audio technology, law enforcement authorities and artificial intelligence experts have expressed concerns regarding its potential exploitation to automate cyberattacks. The first reported case of financial fraud using deep fake technology was published on The Wall Street Journal back in March 2019, where the chief executive officer (CEO) of a British energy firm made a transfer of $243,000 to the swindler. According to the company's insurance firm, Euler Hermes Group SA, the caller demanded an urgent request to send the funds to the account of a certain Hungary supplier. The anonymous CEO testified that he believed himself to be speaking on the phone with his boss from the firm's German parent company and made the transfer as he recognized the boss' "slight German accent" and the "melody of his voice on the phone" – all of which were highly congruous the boss' usual speaking voice and tone.

Though phone scams – also known as voice phishing – are nothing new, the emergence of deep fake audio technology makes it more reliable as attackers could easily use publicly available audio files from social media platforms to disguises as one's relative, friend or co-worker like the case mentioned earlier. The sense of familiarity created by AI decreases one's awareness and ability to detect the mal intent behind the phone, which makes it easy for the attackers to obtain sensitive information and defraud money. Currently, it is not a major concern as not many cases as such were reported since the initial incident, however, experts both within the field of cybercrime and artificial intelligence predict that the number is expected to grow.

## *Political Manipulation*

Another avenue for the potential abuse of deep fake technology lies within politics. It is expected

that the greatest threat will arise from the misuse of the technology in political disinformation campaigns. For example, suppose that someone intentionally leaked a digitally manipulated video of a certain candidate making comments that were politically inflammatory and inappropriate approaching the election season. The proliferation of such media online would then have a profound, irreversible impact on the image of the victim and the political party in which they belong to, leading to swayed election results thereby threatening democracy and destabilizing governments.

Regarding this issue, Professor Hao Li of the University of Southern California noted, “Elections are already being manipulated with fake news, so imagine what would happen if you added sophisticated deep fakes to the mix?” Though there has not yet been a large-scale incident where deep fake technology was used by a political entity to deliberately sabotage their opponent, the increasingly indistinguishable deep fakes could be misused as a powerful political weapon in the future, not only within a nation but also between nations, if not properly regulated.

## International Actions

### *Deep Fake Detection Challenge (DFDC)*

In September 2019, Facebook partnered with industry leaders and experts – Microsoft and Amazon Web Services (AWS) – to launch an open, cooperative Deep Fake Detection Challenge (DFDC) in hoping to expedite research in innovative and effective methods to identify deep fake videos online. Facebook provided the participants with new dataset which consisted of 124,000 videos featuring 3,500 actors and 38.5 days’ worth of data: all of which were collected with consent from paid actors for the sole purpose of research on deep fake media.

*An advert for the Fake Detection Challenge (DFDX)*

The competition provided a platform for 2,114 experts around the world to come together to present and compare their own models with the others thereby allowing them to benchmark their deep fake detection models, brainstorm new methods, and most importantly learn from their peers. A public leaderboard was available where the participants were able to track real time on which of the models proved to be most successful (in terms of accuracy in %) when trained and tested with the dataset provided. In the end, the best-performing models managed to achieve 82.56% accuracy on the public dataset and 65.18% against the black box dataset. Facebook AI mentioned in an article published in June

2020 that "we believe that challenges and shared datasets are key to faster progress in AI." The research community aims to take the outcomes of the challenges further by identifying common, or unique, characteristics of each of the successful models to build the most effective detection technique to protect the society at large.

## Key Players

### *Twitter*

Twitter is proactively involved in order to prevent disinformation from proliferating online by flagging the tweets that does or could potentially contain deep fakes. When a user is viewing tweets that contain such synthetic data and attempts to retweet, like, or comment, they are immediately notified with a warning sign. Twitter also has the right to delete any tweets that contain deep fakes or doctored data if these are thought to be harmful in anyway, which furthers ensures the prevention of unintended spread of individuals' information. In the coming years, Twitter is also planning to provide users with a link that leads onto a credible source of article or video relating to the topics being discussed in order to take actions against disinformation. In 2020, Twitter has publicly asked users for their interest in partnering with them to work together to develop deep fake detection technologies.


*An online user survey conducted by Twitter regarding responses to deep fake media*

### *Facebook*

As being the most widely used social media platform in the world, Facebook has fully embraced their responsibility and made efforts towards the development of deep fake detection tools by hosting Deep Fake Detection Challenge, as mentioned earlier. Facebook has also announced the change in their policy regarding the regulation of deep fake contents on their platform. The new policy states that they will "ban intentionally misleading deep fakes from its platform" and Adam Schiff – the Chairman of the House Permanent Select Committee on Intelligence – responded that it is a "sensible" and "responsible" step.

# Possible Solutions

## *Strengthening Security at Work*

*An office telecommmunication system*

In order to prevent cyberattacks targeting business processes, it is most important to identify the places within the security system that are most susceptible under attack. Therefore, it is vital that cybersecurity systems of firms and companies are regularly examined to ensure minimal chance of success for an external invasion and theft of information using AI-based deep fake technologies. The solutions to combating deep fakes do not always have to be costly or technological. It could be as simple as setting up an office telecommunication infrastructure that is not available to others unless given access or keeping biometric security features up to date. Also, educating the employees of the potential deep fake cyberattacks would be largely effective as such crimes seem to target and deceive people, not computers. Perhaps, a company may consider introducing semantic passwords for conversations to make sure that the person on the phone is the intended caller, not just vocally resembling.

## *Establishing Legal Measures*

Currently, not every country around the world have legal measures in place that efficiently address the problems with the fabrication and distribution of sensitive and/or inappropriate deep fake media online, leaving individuals unprotected from potential harm and encouraging the proliferation of sensitive media. However, introducing substantial penalties as a consequence of producing and sharing deep fake contents will act as a hindrance and discourage misbehavior from the users online. Therefore, nations should prioritize the implementation of a through and effective policy for a safer and cleaner use of deep fake technology online.

# Glossary

## *Artificial Intelligence*

The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

*Deep Learning*

An artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for use in decision making.

*Generative Adversarial Network (GAN)*

A deep neural network framework which can learn from a set of training data and generate new data with the same characteristics as the training data.

*Machine Learning (ML)*

A concept that a computer program can learn and adapt to new data without human interference. Machine learning is a field of artificial intelligence.

*Manipulate*

Alter, edit, or move text or data on a computer.

*Media Literacy*

The ability to access, analyze, evaluate, and understand media sources.

*Neural Network*

a computer system modeled on the human brain and nervous system

## Sources

Arnold. “Deepfake History: When Was Deepfake Technology Invented?” *Deepfake Now*, 9 Jan. 2021, deepfakenow.com/deepfake-history-when-invented/.

Arnold. “The History of Deepfake Technology: How Did Deepfakes Start?” *Deepfake Now*, 21 Apr. 2020, deepfakenow.com/history-deepfake-technology-how-deepfakes-started/.

“Deepfake Detection Challenge Dataset.” *Facebook AI*, 25 June 2020, ai.facebook.com/datasets/dfdc/.

“Deepfake Detection Challenge Results: An Open Initiative to Advance AI.” *Facebook AI*, 12 June 2020, ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.

Devin Partida Jul 22. “Deepfakes Are a Problem, What's the Solution?” *AT&T Cybersecurity*, 5 Aug. 2019, cybersecurity.att.com/blogs/security-essentials/deepfakes-are-a-problem-whats-the-solution.

**Harris, Laurie A. “Deep Fakes and National Security.” Congressional Research Service, 8 June 2021.**

Levin, Adam. “Will Deepfakes Be a Cyber Threat in 2021?” *CyberScout*, 18 Dec. 2020, cyberscout.com/en/blog/will-deepfakes-be-a-cyber-threat-in-2021.

“Rep. Schiff Statement on Facebook's Deepfake Policy.” *Home*, 7 Jan. 2020, schiff.house.gov/news/press-releases/rep-schiff-statement-on-facebooks-deepfake-policy.

**Somers, Meredith. “Deepfakes, Explained.” *MIT Sloan*, 21 July 2020, mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained.**

Stanton, Charlotte. “How Should Countries Tackle Deepfakes?” *Carnegie Endowment for International Peace*, 28 Jan. 2019, carnegieendowment.org/2019/01/28/how-should-countries-tackle-deepfakes-pub-78221.

Stupp, Catherine. “Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case.” *The Wall Street Journal*, Dow Jones & Company, 30 Aug. 2019, www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.