

**Kaunas University of Technology**

Faculty of Informatics

The Dangers of “Deepfake” Technology

Individual Work 1 Report

**Prepared by:**

Gustas Klevinskas

Aidas Jankauskas

Domantas Sabaliauskas

Viktoras Dechtiar

**Academic group**

IFF-8/7

Kaunas, 2019-10-08

Table of Contents

[Memorandum 3](#_Toc21454951)

[Abstract summary 3](#_Toc21454952)

[Introduction 4](#_Toc21454953)

[Main body 4](#_Toc21454954)

[Problems 4](#_Toc21454955)

[An example of the power of deepfake technology 4](#_Toc21454956)

[Spotting deepfakes 5](#_Toc21454957)

[Our research regarding deepfakes and fake news 6](#_Toc21454958)

[Conclusion 8](#_Toc21454959)

[Recommendations 8](#_Toc21454960)

[References 9](#_Toc21454961)

# Memorandum

The purpose of this report is to introduce the reader to a phenomenon dubbed “deepfake”, analyze the effects of this technology to our society, analyze our research results and provide possible solutions.

“Deepfake” is a less scientific term to describe a machine learning algorithm that superimposes a selected face onto an existing video clip or generating audio of a person’s voice. This technology has the potential to jeopardize the personal privacy of individuals, steal their identity, or be used as blackmail. Algorithms are being developed in order to mitigate this problem, also there are other solutions, such as banning the use of deepfake technology, or increasing the population’s media literacy and general skepticism of the things we see online.

This report cites information from reputable news sites such as The Wall Street Journal, Fox News and from more technology focused websites: Vice, Interesting Engineering.

# Abstract summary

Humans have made significant leaps forward with AI and machine learning technology. It has advanced enough that ethical problems are starting to arise. Deepfake, a portmanteau of “deep learning” and “fake”, has introduced some concerning dangers. This technology can be used as a means of blackmailing other people. You can steal their identity (their face and voice) and make it look like somebody is saying or doing things they would never do. This in itself is quite frightening, however, it is important to point out that this technology is freely available to use to all individuals, whether sophisticated or not. In our research, we found that less than half of the people that we questioned even were aware that such technology exists. However, we cannot draw meaningful conclusions from our survey – only 25 people were questioned and all of them were KTU students from the IT faculty. Hence our data could be quite distorted.

Solutions for this problem include banning the use of “deepfake” technology (such ban, sadly, could also be utilized as a means of censorship), increasing the media literacy of the general public and creating algorithms that detect whether a video has been edited. The last option is already being worked on.

# Introduction

Destructive lies are nothing new, however the capacity to twist reality has taken an exponential leap forward with “deepfake” technology. Deepfake is a portmanteau of “deep learning” and “fake”. It refers to AI software that can superimpose a digital composite face on to an existing video of a person or simulate another person’s voice. This capability makes it possible to create audio and video of real people saying and doing things they never said or did. Machine learning techniques are escalating the technology’s sophistication, making deep fakes even more realistic and progressively impervious to discovery. “Soon, it’s going to get to the point where there is no way that we can actually detect deepfakes anymore, so we have to look at other types of solutions,” says the University of Southern California professor Hao Li. Deepfake technology in its nature is relatively easy to acquire and utilize in nefarious ways for both sophisticated and unsophisticated individuals.

The term first rose to prominence when Motherboard reported on a Reddit user who was utilizing AI to superimpose the faces of film stars on to existing adult videos, creating (with varying degrees of realness) pornographic content starring Emma Watson, Gal Gadot, Scarlett Johansson and an array of other female celebrities.

The term deepfake itself comes from the handle of a Reddit user – Deepfakes – who made these types of videos and started the /r/deepfakes subreddit to share them.

# Main body

## Problems

As with all new technology, deepfake creation software is only getting more and more accessible. Its applications are also expanding in multiple directions, including producing full body deepfakes, creating real-time impersonations and seamlessly removing elements from videos. Thankfully, the AI is not yet sophisticated enough to produce perfect results without human intervention. Everyday people will be able to create manipulated images and videos that seem “perfectly real” in “half a year to a year,” says industry pioneer Hao Li. Concern is growing worldwide about the negative impacts that deepfakes could have on individuals, communities, and democracies.

While deep-fake technology will bring certain merits, it also will introduce many harms. Both online and physical media already suffers from truth decay, making it interact in toxic ways with our cognitive biases. Deep fakes will exacerbate this problem significantly by enabling corrupt individuals to create such fake media with mind-boggling speed. Both people and businesses will face novel forms of blackmail, exploitation, intimidation, and personal sabotage. The risks to our democracy and to national security are profound as well. A skillfully produced deepfake that is impersonating a CEO of a multibillion-dollar company could send the company’s stocks plummeting. An impersonation of president of a powerful country could also have devastating effects.

## An example of the power of deepfake technology

There are many examples online of a well-executed fraud using AI generated videos or voices. One of the most recent instances is an impersonation of a U.K.-based energy firm’s parent company’s CEO. The CEO of a U.K.-based energy firm thought he was speaking on the phone with his boss, the chief executive of the firm's German parent company, who asked him to send a copious amount of money to a Hungarian supplier. The impersonator said the request was urgent, instructing the executive to pay within an hour, according to the company's insurance firm, Euler Hermes Group. Law enforcement authorities and artificial intelligence experts deducted that criminals used AI to automate the cyberattack. Whoever was behind this incident appears to have used AI-based software to successfully mimic the German executive's voice by phone. The U.K. CEO recognized his boss' slight German accent and the melody of his voice on the phone, said Rudiger Kirsch, a fraud expert at Euler Hermes, a subsidiary of Munich-based financial services company Allianz (The Wall Street Journal, 2019).

Many governments are grappling for the best way to deal with online misinformation. But some activists and scholars are wary of an outright ban of deepfake technology. They worry that if a law gives government officials the power to decide what is true or false, there is a risk that it might be used to censor unpopular or dissenting views. Hao Li defends the development of deepfake technology, saying the real issue is flagging those with manipulative intent.

## Spotting deepfakes

If you’re worried about the malevolent potential of a deepfake video, you are not alone – so is Facebook. The company has launched a project to sniff out deepfake videos, and it’s pledging more than $10m to the cause. It has pulled in a range of partners including Microsoft for help.

AI relies on lots of data to generate its images, so to create AI that spots deepfakes, Facebook must come up with its own dataset. It will take unmodified, non-deepfake videos, and then use a variety of AI techniques to tamper with them. It will make this entire dataset available to researchers, who can use it to train AI algorithms that spot deepfakes.

In the category of technical solutions, many platforms, researchers and startups are exploring using AI to detect and eliminate deepfakes. There are also new innovations in video forensics that aim to improve our ability to track the authenticity of images and videos, such as ProofMode and TruePic, which aim to help journalists and individuals validate and self-authenticate media (Guardian Project, 2017).

This dataset will hopefully aid in the detection of AI generated audio and video content. In June 2019, researchers at the University of Southern California’s Information Sciences Institute created a model to detect inconsistencies in motion that lead to strange facial movements (Vice, 2019). The University at Albany is also looking for a lack of blinking as many deepfakes often do not blink (The Conversation, 2018).

There are already plenty of widely shared videos that use regular video editing (and not deepfake technology) to disseminate misinformation. Producing such content is not as fast as generating them with AI technology, and in the future it will most likely be outperformed by AI videos in terms of realism. However, if you believe in the message that is being presented anyway, you are not looking for signs that would indicate that the video is not genuine.

Because of this, another solution to deepfake videos and audio recordings needs to be found other than detection algorithms. It involves, as others have pointed out, increasing media literacy among vast swathes of the population, so that they are able to spot fake news when they see it. Sadly, this may be more difficult to achieve than creating a good detection algorithm.

We can expect both AI factions to compete in a kind of arms race, with one side creating increasingly convincing videos that could be used for malicious intent and the other side creating AI to detect them. In that scenario, the people who try to detect fake media need all the help they can get.

## Our research regarding deepfakes and fake news

During the time period between September 16th and September 27th we were performing a survey, asking 25 Kaunas Technology University students some questions concerning deepfakes. We have asked questions that, in our opinion, might represent students’ media literacy regarding this technology.

First, we showed a video that was made using deepfake technology and asked students if they saw something wrong in that video.

Figure 1. The results of our first question

Results indicate that most university students cannot distinguish a deepfake video from a real video. Thankfully, 44% of respondents were attentive and skeptical enough to spot that something was amiss in the video.

Secondly, we have asked if students were aware of deepfake technology itself.

Figure 2. The results of our second question

Results show that many students still don’t know about this technology or are only aware vaguely. The third question that we have asked was how could deepfake technology be utilized.

Figure 3. The results of our third question

Many of respondents haven’t perceived a dangerous use of this technology. This might be because deepfakes are commonly used for internet jokes.

Finally, respondents that previously stated that they have seen a deepfake video before and were aware of this technology, were given a question on how often they see similar media.

Figure 4. The results of our fourth question

Majority of students stated that they encounter this kind of media quite often.

Overall, the results are quite disappointing and indicate that society is extremely vulnerable because even university students are unable to identify deepfake videos, though they encounter them quite often. We could assume that students from non-technology faculties (such as humanitarian) would know about deepfakes even less. It is even more frightening that university students are quite media literate compared to majority of society that we live in.

# Conclusion

With the ever-growing capabilities of AI technology, we must stay more alert to fake media than ever before. As Barack Obama once said: “We’re entering an era in which our enemies can make it look like anyone is saying anything, at any point in time. Even if we would never say those things.”

For now, banning deepfakes, especially as a tool for revenge pornography, is the only recourse. We’ll likely start hearing about new regulations and laws in the near future. Hopefully, programmers will find positive uses for the AI and machine learning before that happens (Fox News, 2019).

However, the advancements in AI technology may also come with some merits. Whilst giving political speeches is not likely to be eliminated by this technology, it may increase the overall skepticism of our community of the things we come across online. That, in turn, would hinder the spread and effect of fake news in general.

As others have noted, we should thank deepfakes for “making us realize once again that we can’t take everything we see and hear for granted. For creating a problem for us to solve, early on, before it becomes so big, and has influenced so many of us incorrectly, that it’s too late.” (Interesting Engineering, 2019).

Now, as for our survey, the results are quite obvious – not that many people are aware that such machine learning technology exists. Also, in *figure 3* it is seen that a large proportion of the students we questioned do not take this technology seriously and believe that the main purpose is creating funny internet videos. It is also important to mention that this survey is quite limited in terms of the conclusions that we can draw. Mainly students from the IT faculty were interviewed, thus making the results strongly biased. Had we had more respondents and a larger variety of them, the results could become drastically different.

# Recommendations

In order to protect yourself from deepfake videos or audio recordings (like the story with the impersonation of a CEO), you must exercise great caution when viewing important content. A regular individual will not have the technical proficiencies in creating deepfake detection algorithms, so for such people, being media literate is the only solution. Just like the story with the impersonation of a CEO, you must not trust people that you have not verified that are who they say they are; reading articles only from recognized and trustworthy sources, watching TV channels that focus not on popular media, but instead are thorough with their research and have a reputation for being reliable.

Also, media literacy ought to be taught in schools, as kids and teenagers are one of the naivest and easily persuaded group of individuals. We, as humans, carry our habits created from early childhood into adulthood. Therefore, educating the young is of utmost importance.

# References

Chesney, R. & Citron, D. K., 2018. *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,* s.l.: s.n.

Fox News, 2019. *Terrifying high-tech porn: Creepy 'deepfake' videos are on the rise.* [Online]   
Available at: https://www.foxnews.com/tech/terrifying-high-tech-porn-creepy-deepfake-videos-are-on-the-rise

Guardian Project, 2017. *Combating “Fake News” With a Smartphone “Proof Mode”.* [Online]   
Available at: https://guardianproject.info/2017/02/24/combating-fake-news-with-a-smartphone-proof-mode/

Interesting Engineering, 2019. *Deepfakes Are Bad but What Are Some of the Possible Advantages?.* [Online]   
Available at: https://interestingengineering.com/deepfakes-are-bad-but-what-are-some-of-the-possible-advantages

The Conversation, 2018. *Detecting ‘deepfake’ videos in the blink of an eye.* [Online]   
Available at: https://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072

The Wall Street Journal, 2019. *Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case.* [Online]   
Available at: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Vice, 2019. *AI Can Now Detect Deepfakes by Looking for Weird Facial Movements.* [Online]   
Available at: https://www.vice.com/en\_us/article/evy8ee/ai-can-now-detect-deepfakes-by-looking-for-weird-facial-movements