TALLINN UNIVERSITY OF TECHNOLOGY
School of Business and Governance
Department of Law

Ibrahim Mammadzada

DEEPFAKES AND FREEDOM OF EXPRESSION: EUROPEAN PERSPECTIVE

Master's thesis
Programme MA in Law, specialization Law & Technology

Supervisor: Agnes Kasper PhD

Tallinn 2021

I hereby declare that I have compiled the thesis independently
and all works, important standpoints and data by other authors
have been properly referenced and the same paper
has not been previously presented for grading.
The document length is 17192 words from the introduction to the end of the conclusion.

Ibrahim Mammadzada ……………………………
(signature, date)
Student code: 195664HAJM
Student e-mail address: ibmamm@ttu.ee

Supervisor: Agnes Kasper PhD:
The paper conforms to requirements in force

……………………………………………
(signature, date)

Chairman of the Defence Committee:
Permitted to the defence
…………………………………
(name, signature, date)

TABLE OF CONTENTS

ABSTRACT

Regulation of AI-powered deepfakes is one of the most recent problems modern democracies face. Because of its close relationship with freedom of expression, the regulation of deepfake technology might have a chilling effect on this human right. Therefore, this thesis assesses interference with free expression in the context of deepfake threats from the perspective of Article 10(2) of the European Convention For Protection of Human Rights and Fundamental Freedoms as well as the relevant case law of the European Court of Human Rights. This research examined article 10 related cases of the court where "necessity" for interference were justified on different grounds. After that, interference in the context of deepfakes was compared to those cases in order to determine when it could satisfy the "necessary in a democratic society" test. The findings of this research support the hypothesis that deepfake threats could create "pressing social need" for intervention on the grounds of protection of national security, territorial integrity, public order, rights and reputation of others, and maintaining impartiality and authority of judiciary. The thesis suggests that the most proportional legal solution to the problem is the enactment of criminal provisions dealing with malicious deepfakes. Such provisions should be designed with maximum precision and contain a wide range of sanctions in order to allow national courts to choose the most "proportionate" one depending on each case. Overall, to effectively combat malicious deepfakes, and protect free expression, legal and technological solutions, as well as awareness-raising in public, should be used combined.

Keywords: Freedom of Expression, Deepfake, Human Rights, ECHR, ECtHR.

INTRODUCTION AND JUSTIFICATION OF THE TOPIC

With its debut in 2017, the deepfake technology caused alarm and panic in democratic societies because of its potential threats to fundamental elements of democracies.[1] And some scholars have already tried to address the challenges that deepfakes pose to individuals and the public as a whole.[2]

But there is not sufficient real-life evidence suggesting that deepfakes have caused significant damage to democratic discourse. Many scholars estimate or construct different scenarios in which deepfakes could be used to undermine democracies. Scholars Danielle Citron and Robert Chesney suggest that "a well-timed and thoughtfully scripted deep fake or series of deep fakes could tip an election, spark violence in a city primed for civil unrest, bolster insurgent narratives about an enemy's supposed atrocities, or exacerbate political divisions in a society."[3] It was anticipated that the deepfake would damage the 2020 US elections in the same way fake news did in the 2016 election.[4] But contrary to suggestions, deepfakes have not been used as a manipulation tactic in the 2020 US Presidential Election; at least, there is no evidence backing it.[5]

However, considering the prevalence of disinformation campaigns[6] used for manipulation of public opinion, it would not be wrong to think that at some point, they will be used for this purpose. Nevertheless, regulating the technology where there is no real-life evidence of any public damage to date is very tricky, especially when this regulation would bring limitations to freedom of expression. Hence before rushing to regulate deepfake technology, its potential threats should be assessed in light of Article 10(2) of the European Convention For Protection of Human Rights and

[1] LaMonaca, J. P. (2020). break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes. American University Law Review, 69(6), 1945-1988. p. 1949

[2] ibid

[3] Chesney, R., & Citron, D. (2018). *Disinformation on Steroids: The Threat of Deep Fakes*. Council on Foreign Relations. Retrieved 11 April 2021, from https://www.cfr.org/report/deep-fake-disinformation-steroids.

[4] Meneses, J. P. (2021). Deepfakes and the 2020 US elections: what (did not) happen. arXiv preprint arXiv:2101.09092. 1-13, p 2

[5] Ibid 7

[6] A report published by the Freedom House found that in 2016 online manipulation and disinformation tactics played an important role in at least 18 national elections that year, including the United States. Freedom House. (2017). Manipulating Social Media to Undermine Democracy. Retrieved from Freedom on the Net 2017: https://freedomhouse.org/report/freedom-net/freedom-net-2017

Fundamental Freedoms(hereinafter ECHR)[7] as well as the relevant case law of the European Court of Human Rights(Hereinafter ECtHR).

As opposed to First Amendment governing the free speech in the US that is written in absolute terms and does give only narrow possibilities for interference, ECHR article 10 paragraph 2 provides an exhaustive list of grounds on which freedom of expression might be limited. At first sight, It seems enough for regulation of deepfakes, but in fact, ECtHR applies strict scrutiny over any type of restriction on any kind of expression. As opposed to the US, where the First Amendment and the supreme court conceptualized the nature of truth and the free flow of information based on the marketplace of ideas approach, the ECtHR employed a different approach regarding the protection of free expression. Under this approach, the information is expected to be a public good and should be dealt with as such. Furthermore, this right is not absolute and should be balanced with other rights enshrined in the ECHR.[8]

According to ECtHR case law, in order state to be able to interfere with freedom of expression, the interference should be prescribed by law, have a legitimate aim and be "necessary in a democratic society." [9] The last "necessary in a democratic society" test has two criteria in it: "Pressing social need" and "Proportionality". For any regulation of expression, there should be "pressing social need" in a democratic society, and if there is, the interference should be proportional to this need.

Because of the newness of deepfake technology, scholarly literature discussing the threats and possible solutions of this matter is sparse.[10] Furthermore, there is no academic literature discussing when and under which circumstances deep fakes can pass the "pressing social need" test put forward by ECtHR. Besides, existing literature tends to locate deepfakes within traditional fake news, even though some scholars suggest that Deepfakes pose a more significant threat than "traditional" fake news because it is more difficult to detect them,[11] and society is inclined to look

[7] Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 November 1950, ETS 5, available at: https://www.refworld.org/docid/3ae6b3b04.html [accessed 12 April 2021]

[8] Schroeder, J. (2020). Free expression rationales and the problem of deepfakes within the EU and US legal systems. Syracuse Law Review, 70(4), 1171-1204. p.1194

[9] Bychawska-Siniarska, D. (2017). Protecting the right to freedom of expression under the European convention on human rights: A handbook for legal practitioners. Council of Europe. p.32

[10] Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11). 40-53. p. 40

[11] Ibid 42

at the facts with the motto of "seeing is believing".[12] Hence, assuming deepfakes are just a form of fake news, and every research and regulation on fake news could be applied to this phenomenon as well, has the potential to be wrong. Deepfake technology and its threats deserve separate and detailed research.

Moreover, an all-out ban of Deepfakes would be a direct violation of freedom of expression since it has many benign applications such as advertising, educational, artistic, satire or parody. This also gives rise to grave concerns that malevolent deepfake users can cause immense damage and then wrap themselves "in the protective cloak of free expression."[13]

Taking into account the aforementioned factors, this research will evolve around the hypothesis that the threats of Deepfakes create "pressing social need" for intervention. Therefore, in light of this hypothesis, the research question of this thesis will be as follows:

When does the interference with free expression(art.10) in the context of deepfakes pass the "necessary in a democratic society" test put forward by the ECtHR?

As regards the methods, this research is conducted using doctrinal analysis methods to systematically study ECtHR's case law. With a doctrinal approach, the thesis identifies the main patterns of ECtHR's application of the "necessary in a democratic society" test. To that end, this research examined article 10 related cases of the court where "necessity" for interference were justified on different grounds. After that, interference in the context of deepfakes was compared to those cases in order to determine when it could satisfy the "necessary in a democratic society" test. Cases were taken from HUDOC, which is an official database of ECtHR. To find relevant cases, I have used the following keywords:

Freedom of Expression, National security, Territorial integrity, Necessary in a democratic society, proportionality, pressing social need, public order, reputation and rights of others, fake news.

For answering the abovementioned research questions, this thesis will follow the following structure. The first chapter will present information on the technological background, and then chapter two will identify benign and malicious uses of deepfakes. The third chapter presents a detailed discussion of freedom of expression perspective of ECHR, and the three-part test put

[12] Pfefferkorn, R. (2020). "deepfakes" in the courtroom. Boston University Public Interest Law Journal, 29(2), 245-276.p. 256

[13] Mostert, F. (2020). 'Digital due process': A need for online justice. *Journal of Intellectual Property Law & Practice, 15*(5), 378-389. p.382

forward by ECtHR to assess violations of article 10. Chapter 4 will proceed to the detailed examination of the cases and identifying how the court determines the existence of “pressing social need” on different grounds. Then the court’s practice on the identification of “pressing social need” will be applied to threats posed by deep fakes. Chapter 5 will examine different interference methods to identify which method is more likely to pass the “proportionality” test in the context of deepfake threats.

1. TECHNOLOGICAL BACKGROUND

The technology behind deep-fakes is called Generative Adversarial Networks (hereinafter - GANs technology) and was first introduced in 2014 by Ian Goodfellow.[14] It should be noted that the fabricated audiovisual material existed even before the GANs technology, but it improved the fidelity of these fabrications so drastically that, sometimes, it requires sophisticated forensic tools to identify whether given audiovisual material is fake or real.[15]

GANs technology is composed of two distinct artificial neural networks that work against each other. The first generative neural network creates fake datasets, and the other discriminator network, with the help of real datasets, identifies the fake data.[16] This process continues, and with time generative network creates such sophisticated fake data that it becomes harder and harder for discriminator system to identify.[17] The race between these two neural networks is in constant escalation, as each learns the methods of the other, resulting in improvement of the ability to create data samples with increasing fidelity.[18]

Although Generative models based on deep learning are prevalent, GANs are one of the most successful generative models, especially considering their ability to create high-resolution fake images.[19] According to Goodfellow et al.," The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles."[20]

[14] Pfefferkorn (2020) supra nota 12. 247

[15] LaMonaca (2020) supra nota 1. 1952

[16] Pfefferkorn (2020) supra nota 12. 247

[17] ibid.

[18] Nicholson, Chris. "A Beginner's Guide To Generative Adversarial Networks (Gans)". *Pathmind*, https://wiki.pathmind.com/generative-adversarial-network-gan. Accessed 26 Mar 2021.

[19] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144. p. 139

[20] ibid

1.2. DIFFUSION OF DEEPFAKES

However, although first introduced in 2014, the first widespread use of this technology has begun in 2017 when a Reddit user nicknamed “Deep-Fake” used this technology to create fake pornographic videos of celebrities.[21] The technology quickly drew attention resulting in a specifically created subreddit, “r/deepfakes,” where users were posting pornographic videos of face-swabbed celebrities. This also caused digital impersonations to be labeled as "deepfake" from that point on, and the word is used as an umbrella term to cover a full range of realistic digital falsification of image, video, and audio materials.[22]

The “r/deepfake” subreddit also contained the script describing the procedure of deepfake creation as well as the latest know-hows.[23] In January 2018, another Reddit user created the FakeApp to make it available for everyone, including those who do not have the technical skills.[24] Within two months, the number of forum members had reached 25,000, rising concerns. Reddit taking into account the concerns, decided to shut down the forum. [25] But it did not stop the development and democratization of deepfakes, as there are still over 20 communities and online forums dedicated to deepfake creation. 13 deepfake communities on the internet have approximately 100,000 users combined.[26] Furthermore, deepfake programs are proliferating with drastic speed and becoming available to everyone, including those with malicious intent. "Reface," the deepfake video creation app, has been downloaded more than 70 million times since it was first launched in January 2020. It is also one of the ‘top five’ most popular apps in around 100 countries.[27]

[21] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, *1*. p.1

[22] Chesney, B., & Citron, D. (2019). Deep fakes: looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753-1820. p. 1757

[23] Panyatham, Paengsuda (2021) "Deepfake Technology In The Entertainment Industry: Potential Limitations And Protections — AMT Lab @ CMU". *AMT Lab @ CMU*, https://amt-lab.org/blog/2020/3/deepfake-technology-in-the-entertainment-industry-potential-limitations-and-protections.

[24] Meškys, L., Liaudanskas, G., Kalpokienė, J., & Jurcys, P. (2020). Regulating deep fakes: legal and ethical considerations. *Journal of intellectual property law & practice. Oxford: Oxford university press, 2020, vol. 15, iss. 1*. p.24

[25] ibid

[26] Panyatham(2021) supra nota 23.

[27] Lomas, N. (2020, December 8). *Reface grabs $5.5M seed led by A16z to stoke its viral face-swap video app*. Techcrunch. https://techcrunch.com/2020/12/08/reface-grabs-5-5m-seed-led-by-a16z-to-stoke-its-viral-face-swap-video-app/

Although 96 % of existing AI-powered Deepfakes are of pornographic nature,[28] the technology could be applied in a wide range of areas, be it with malicious or benign intent. The next chapter will look at the different applications of this technology.

[28] Romano, Aja. (2019) "Deepfakes Are A Real Political Threat. For Now, Though, They'Re Mainly Used To Degrade Women.". *Vox*, https://www.vox.com/2019/10/7/20902215/deepfakes-usage-youtube-2019-deeptrace-research-report. Accessed 1 Mar 2021.

2. HOW IS IT USED?

2.1 BENEFICIAL USE OF DEEPFAKES

Deepfake Technology, although caused significant fear, has a wide variety of beneficial uses such as artistic, Research, educational, and a way of expression of opinions.[29] Exploring the beneficial use of deepfakes and the underlying technology is important, as it helps to understand the relationship between freedom of expression and deepfakes and the reason why it is difficult to tailor a regulation of this technology without damaging human rights.

2.1.1. ARTISTIC USES

Disney Research Studios, by using progressive algorithm training, stabilization technology, and lighting effects, managed to create hyper-realistic face swabs at a megapixel resolution.[30] This research is of great importance considering the fact that, so far, deepfake was not used in filmmaking, as its fidelity at high resolutions was very low. But this research will, no doubt, pave the way for expanding the use of the technology for commercial projects. Although in filmmaking different face swabbing methods have been used so far, they are "typically elaborate and labor-intensive computer-graphics methods" and require "extensive frame-by-frame animation and post-processing by digital-effects professionals."[31] Furthermore, these methods are very expensive and time-consuming to the point where sometimes it takes months to produce even a mere second of face-swabbed footage.[32]

Deepfake technology can be used to reverse aging and create a younger version of an actor or to edit scenes when it is not possible for an actor to participate.[33] For example, after the death of actor Paul Walker, his scenes have been edited by the abovementioned computer graphics methods to finish the movie.[34]

[29] Chesney & Citron (2019) supra nota 22. 1763

[30] Naruniec, J., Helminger, L., Schroers, C., & Weber, R. (2020). High-Resolution Neural Face Swapping for Visual Effects. *Computer Graphics Forum, 39*(4), 173-184. p. 182

[31] Ibid. 2

[32] ibid

[33] Panyatham(2021) supra nota 23

[34] Caulfield, AJ. (2020) "The Truth About Recreating Paul Walker For Fast And The Furious - Exclusive". *Looper.Com*, https://www.looper.com/184468/the-truth-about-recreating-paul-walker-for-fast-and-the-furious/. Accessed 16 Mar 2021.

Another example of artistic use of Deepfake technology is the "Dalí Lives" project. Dali Museum, in cooperation with the advertising firm Goodby Silverstein & Partners of San Francisco (GS&P), created the likeness of Salvador Dali using cutting-edge technology. The "Dali Lives" project gives visitors an opportunity to interact "with an engaging, lifelike Salvador Dalí on a series of screens throughout the Museum."[35] At the end of the tour, Dali even asks the visitor whether he/she wants a selfie and turns his back with the phone in his hand to take a pic which later is sent to the visitor via email.[36]

2.1.2 RESEARCH AND EDUCATION

The Educational and research potential of Deepfake technology cannot be ignored. It can be used in different areas for research and educational purposes. For example, the underlying technology of Deepfakes could help to synthesize realistic data for researchers to develop new diagnostic methods and treatments for diseases without using the actual data of patients, thus protecting their privacy.[37] With this technology, even a single hospital can create "an entirely imaginary population of virtual patients" without sharing real patient data.[38] In 2018, the study conducted by Nvidia, the Mayo Clinic, and the MGH & BWH Center for Clinical Data Science showed promising results in this regard. In this study, with the help of GANs technology, researchers created "fake" MRI scans, which in turn used to train machine learning to spot tumors.[39] In addition, different studies using GANs technology created synthetic medical data for research, detection, and treatment of liver and skin lesions.[40][41]

[35] "Dalí Lives: Museum Brings Artist Back To Life With AI - Salvador Dalí Museum". *Thedali.Org*, 2019, https://thedali.org/press-room/dali-lives-museum-brings-artists-back-to-life-with-ai/.

[36] Kelleher, S. (2021). *AI Has Resurrected Salvador Dali And Now He's Your Museum Tour Guide*. Forbes. Retrieved 12 April 2021, from https://www.forbes.com/sites/suzannerowankelleher/2019/05/13/ai-has-resurrected-salvador-dali-and-now-hes-your-museum-tour-guide/?sh=76c050ad2d97.

[37] Chandler, S. (2020). Why Deepfakes Are A Net Positive For Humanity. Forbes. Retrieved 12 April 2021, from https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity/?sh=6d89f7cc2f84.

[38] Geraint, R. (2019). *Opinion: How the technology behind deepfakes can benefit all of society*. UCL News. Retrieved 12 April 2021, from https://www.ucl.ac.uk/news/2019/nov/opinion-how-technology-behind-deepfakes-can-benefit-all-society.

[39] Shin, H. C., Tenenholtz, N. A., Rogers, J. K., Andriole, K. P., Michalski, M. G., Schwarz, C. L., . . . Gunter, J. (2018). Medical image synthesis for data augmentation and anonymization using generative adversarial networks. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11037*, 1-11. p. 1

[40] Frid-Adar, M., Diamant, I., Greenspan, H., Klang, E., Amitai, M., & Goldberger, J. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing, 321*, 321-331.

[41] Baur C., Albarqouni S., Navab N. (2018) Generating Highly Realistic Images of Skin Lesions with GANs. In: Stoyanov D. et al. (eds) OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy,

Deepfakes can also revolutionize history classes with interactivity which would maximize the learning experience. Historical events can be narrated by the historical figures themselves. A good example of it would be hologrammatic interviews created by the Illinois Holocaust Museum and Education Centre. The museum created an environment where visitors could ask questions to holocaust survivors and listen to their stories.[42]

Edtech company Udacity developed GANs based algorithm called "LumièreNet," which generates synthetic educational videos from audio narrations.[43] It will be very beneficial for mass open online course (MOOC) platforms such as Coursera, as creating video content is very time-consuming and expensive.

Deepfakes can also be used for automatically changing the language of videos, thus cutting the cost of translation and voice recording. This is evident in the video recording of David Beckham, where he gives information about the Malaria Must Die Campaign in 9 languages with his own voice.[44]

2.1.3 SELF-EXPRESSION

Deepfakes by anonymizing voices and faces can help human rights defenders, activists, and journalists who operate under oppressive regimes to express their opinions, disseminate news and stay anonymous. Via deepfakes, individuals may create avatars for online expression. These digital avatars would provide autonomy and privacy to human rights defenders, activists, and journalists, thus helping to extend their purposes, ideas, and beliefs and enable free expression.[45] People suffering from certain diseases such as ALS can benefit from this technology to use their own voice when speaking.[46]

Clinical Image-Based Procedures, and Skin Image Analysis. CARE 2018, CLIP 2018, OR 2.0 2018, ISIC 2018. Lecture Notes in Computer Science, vol 11041. Springer, Cham. https://doi.org/10.1007/978-3-030-01201-4_28 p.267

[42] Braunstein, E. (2018). At this Holocaust museum, you can speak with holograms of survivors. Timesofisrael.com. Retrieved 12 April 2021, from https://www.timesofisrael.com/at-this-holocaust-museum-you-can-speak-with-holograms-of-survivors/.

[43] Kim, B., & Ganapathi, V. (2019). Lumi`ereNet: Lecture Video Synthesis from Audio.

[44] David Beckham Travels to the Future to Announce the End of Malaria - IVCC. IVCC. (2021). Retrieved 12 April 2021, from https://www.ivcc.com/david-beckham-travels-to-the-future-to-announce-the-end-of-malaria.

[45] Chesney & Citron (2019) supra nota 22. 1771

[46] Ibid.

Moreover, deep fakes could be used for parody or satire purposes. And if considered as a medium, it can facilitate political debates and creative interactions. This could inevitably lead deepfakes to become an inseparable part of free expression.[47]

2.2. POTENTIAL THREATS

The threats posed by Deep fakes could be divided into two categories:
1) Individual-level threats which include revenge porn, harassment[48], blackmail, identity theft, and deepfake powered social engineering attacks.[49]
2) Public-level threats, which include dissemination of fake news, manipulation of elections, deepening socio-political polarization, endangering national security, and so on.

2.2.1 INDIVIDUAL LEVEL THREATS

Revenge porn is sexually explicit images that are created and disseminated to humiliate, threaten, exploit or harm a romantic ex-partner.[50] Later this definition was expanded to cover nonconsensual pornographic deep fakes that are disseminated by perpetrators for financial or other unlawful gains.[51] These doctored porno videos are also considered as a form of "image-based sexual abuse," which is a broader term.[52]

Deepfakes add up to the already existing revenge porn problem. Deepfake pornography, when created without consent, amounts to an invasion of privacy.[53] With this technology, now perpetrators do not even have to obtain sexually explicit images of the victim. Using non-explicit images taken from social media, they simply superimpose the faces of victims onto pornographic

[47] Meskys et al (2020) supra nota 24. 29
[48] Harris, D. (2018-2019). Deepfakes: False Pornography Is Here and the Law Cannot Protect You. Duke Law & Technology Review, 17, 99-128. p.102
[49] Chesney & Citron (2019) supra nota 22. 1772
[50] Citron, D. K., & Franks, M. A. (2014). Criminalizing revenge porn. *Wake Forest L. Rev.*, *49*, 345-392. p.346
[51] Hayward, P., & Rahn, A. (2015). Opening Pandora's Box: pleasure, consent and consequence in the production and circulation of celebrity sex videos. *Porn Studies*, *2*(1), 49-61. https://doi.org/10.1080/23268743.2014.984951
[52] McGlynn, C., & Rackley, E. (2017). Image-Based Sexual Abuse. *Oxford Journal Of Legal Studies*, *37*(3), 534-561. https://doi.org/10.1093/ojls/gqw033 p.540
[53] Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, *7*(4), 415-423. https://doi.org/10.1080/23268743.2020.1757499. p.416

video and disseminate or threaten to disseminate it.[54] Identifying the disseminator is very difficult, and If the deepfake is shared online, the delay between taking legal action and obtaining an injunction will enable the deepfake to circulate and do immense damage. Instant delivery of content and relatively slow legal proceedings make the deepfake an ideal crime tool for malevolent actors.[55]

As a way of extortion, blackmailers might use deep fakes to force victims to provide something of value such as money, commercial secrets, or nude images(sextortion) under the threat of dissemination.[56] The threat of making the video public could be so powerful that it can force victims to engage in unlawful or violent acts. For example, a person who has confidential information or state secrets might be forced to share the information with foreign intelligence to prevent the release of the video. Also, malevolent actors might extract ransom from the victims by showing deepfake videos depicting the kidnapping of their loved ones.[57]

Furthermore, there is real-life evidence proving that deepfakes are already used as part of social engineering scams. For example, in 2019, deepfakes have been used to defraud The CEO of UK based energy company by convincing him that he was getting orders from the chief executive of the German parent company.[58]

Deepfake can also easily fool facial recognition technologies. Therefore public and private organizations using such technologies which store huge amounts of image data of individuals for verification and security purposes should address the challenges of identity theft in case of data leakage.[59]

[54] Öhman, C. (2019). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology, 22*(2), 133-140. p. 139

[55] Perot, E., & Mostert, F. (2020). Fake it till you make it: an examination of the US and English approaches to persona protection as applied to deepfakes on social media. *Journal of Intellectual Property Law & Practice*, *15*(1), 32-39. p.38

[56] Chesney & Citron (2019) supra nota 22. 1772

[57] ibid

[58] Damiani, J. (2019, September 3). A Voice Deepfake Was Used To Scam A CEO Out Of $243,000. Forbes. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=34f4033c2241

[59] Westerlund (2019) Supra nota 10. 45

2.2.2. PUBLIC LEVEL THREATS.

For now, the cases of deepfake uses other than nonconsensual pornography are relatively low because of its newness. But as the technology matures and improves, different ways of unlawful use are becoming more likely.[60] When it comes to public-level threats, a wide array of possible misuses exist. The specific feature of deepfakes posing public-level threats is that they are part of information warfare, acting as a powerful weapon to boost disinformation campaigns and further blur the line between truths and lies. Below different potential threats of Deepfakes will be examined.

2.2.2.1 Disruption of Democratic Discourse and Polarization of Society

When there are no agreed-upon facts and the public is divided on even the most simple and fundamental matters such as climate change, vaccines, it is extremely difficult to establish and maintain well balanced democratic discourse.[61] When there is no shared truth, and people only believe the things that confirm their preexisting beliefs, it disrupts the democratic discourse and leads to deep polarization of society.

In deeply polarized societies, people who do not share the same opinions and beliefs antagonize each other. The United States is one of the countries that deeply suffers from political divisions. Republicans and Democrats each truly believe and fear that the other will destroy the country.[62] It is evident in the 2020 presidential elections in the US, as Trump supporters stormed the capitol claiming that the election was frauded, although there was not even single evidence supporting their claim. It was an event that has never been experienced in US history before.[63]

[60] Brown, N. I. (2020). Deepfakes and the Weaponization of Disinformation. Virginia Journal of Law & Technology, 23, 1-59. P.9

[61] Post, R. C. (2018). Data privacy and dignitary privacy: Google spain, the right to be forgotten, and the construction of the public sphere. *Duke Law Journal, 67*(5), 981-1072. p.1005

[62] Kleinfeld, Rachel, and Aaron Sobel. *Eu.Usatoday.Com*, 2021, https://eu.usatoday.com/story/opinion/2020/07/23/political-polarization-dangerous-america-heres-how-fight-column/5477711002/. Accessed 21 Mar 2021.

[63] Beaumont, Peter et al.(2021) "How A Mob Of Trump Supporters Stormed The Capitol – Visual Guide". *The Guardian*, https://www.theguardian.com/us-news/2021/jan/07/how-a-mob-of-trump-supporters-stormed-the-capitol-visual-guide. Accessed 26 Mar 2021.

The 2016 US election also has abundant examples of the abovementioned divisions on even the simplest facts, such as the birthplace of President Obama. According to the survey conducted in 2017, 42% of Republicans believed that Barack Obama was not born on US soil, in spite of the fact that there was a substantial amount of evidence debunking it.[64]

Although fake news circulating prior to and during the US Elections did not contain deepfakes, one can easily imagine how profound the impact of deepfaked video supporting above mentioned rumors would have been if they were used.

Another way to use deepfake nefariously is not to use it. In other words, the mere existence of deepfake technology can be used to create doubt and confusion in public.[65] In 2019 a real video of Ali Bongo, the president of Gabon, was claimed to be deepfake, and it is taught that this allegation provoked a military coup attempt against him.[66]

2.2.2.2.The Threat to Public Order

Deepfake video can achieve what mere textual false information can not. For instance, a video depicting police officers shooting an unarmed person while using racial slurs would be a last straw on the rising racial tensions and cause civil unrest or riots. Imagine a video depicting just that released into the public. After all the riots and protests because of the killing of Georg Floyd by the police officer,[67] that would send the furious public to the edge. In Malaysia, a deepfake video of a man confessing his intimate relationship with one of the local cabinet ministers created political controversy.[68]

Deepfake can also undermine public order by creating false panic. For instance, a deepfake based nuclear missile attack rumor or a video depicting the spread of a deadly outbreak could create a false sense of panic and crisis. In 2018, a false ballistic missile attack alert caused mass panic and

[64] Brown (2020) supra nota 60. 10

[65] Smith, Hannah, and Katherine, Mansted.(2020) *Weaponised Deep Fakes: National Security And Democracy*. ASPI International Cyber Policy Centre, 1-21. p. 13

[66] ibid

[67] Aratani, L. (2020). *George Floyd killing: what sparked the protests – and what has been the response?*. the Guardian. Retrieved 11 May 2021, from https://www.theguardian.com/us-news/2020/may/29/george-floyd-killing-protests-police-brutality.

[68] Westerlund (2019) supra nota 10. 44

terrified the population in Hawaii.[69] Although it was a mistake and the warning was retracted 40 minutes later, it forms a good example of how the public is vulnerable in this sense.

It is not difficult to imagine how continuous exposure to deepfake audio or videos depicting officials warning the public about upcoming terrorist attacks, outbreaks, or contamination of drinking water with radio actives could cause tremendous distress and panic. At some point, it could create a situation where people might develop information apathy and ignore even real and truthful warnings.

2.2.2.3. Threats To National Security

Above mentioned threats could also undermine national security. Politically motivated individuals or organizations, or even adversary states, might produce and disseminate deepfake videos depicting elected officials taking bribes, engaging in treacherous espionage activities, uttering racially biased and genocidal words. These kinds of deepfakes could lead to erosion of public trust in authorities and could even sway an election.[70] By doing so, even the existence of the state could be threatened.

Real-life evidence exists in the context of Russian involvement in the 2016 presidential election where US intelligence concluded that Russians conducted “extensive influence operations” to “undermine public faith in the U.S. democratic process, denigrate Secretary Clinton, and harm her electability and potential presidency.”[71]

Michael McFaul, the American ambassador in Russia from 2012-2014, said that Russia had used doctored disinformation videos against him and other politicians for years. He alleged that Russian state propaganda superimposed his face onto videos depicting him saying and doing things he never said or did. He said that these videos were used to accuse him of pedophilia.[72] It is probable that with time deepfakes will be used more often to strengthen the abovementioned nefarious efforts. Elected officials or/and individuals with access to classified information will be at high

[69] "Hawaii Worker Who Sent Missile Alert Was '100% Sure' Attack Was Real". *The Guardian*, 2018, https://www.theguardian.com/us-news/2018/feb/03/hawaii-worker-sent-missile-alert-100-percent-sure-attack-real.

[70] Sayler, Kelley M., and Laurie A. Harris. (2020) *Deep Fakes And National Security*. Congressional Research Service, 1-3, https://crsreports.congress.gov/product/pdf/IF/IF11333. p.1

[71] ibid

[72] Hall, H. (2018). Deepfake videos: When seeing isn't believing. Catholic University Journal of Law and Technology, 27(1), 51-76. p.60

risk of being subject to such deepfake based blackmail attempts.[73] When at war, deepfakes could be used to depict military personnel engaging in war crimes. This would be very powerful to turn the international public opinion against a particular state whose military was depicted in those videos.

Furthermore, these doctored videos might be very beneficial in military operations from strategic and operational perspectives, as they can be used to confuse or deceive the army into believing that it will be attacked from a particular direction, whereas real attacks will come from another.[74] Deepfakes could be used in a myriad of ways to undermine national security. It is a matter of one's imagination how to use them to do so.

2.2.2.4. Threat to judiciary.

Judiciary could suffer from deepfakes the most, as audiovisual material is considered one of the most reliable pieces of evidence. Deepfake will challenge the video evidence admissibility standards. The judiciary will be affected by this technology in two ways:[75]
1) Firstly, deepfakes could be presented as real audiovisual evidence, thus resulting in people convicted falsely.
2) Secondly, benefiting from the uncertainty caused by deepfakes, defendants may challenge the authenticity of real audiovisual material by claiming that it is doctored, thus getting rid of the liability.[76] Scholars Chesney and Citron refer to this phenomenon as "Liar's Dividend."[77]

The first case scenario had occurred in the UK when doctored audio material was presented to the court in a child custody battle. The audio evidence was implying that the husband was making violent threats to his wife. But in fact, the audio was edited to include those "threats".[78] It was the first-ever recorded case of deepfakes used to manipulate the judicial process. But there will be many more if the technology keeps the current rate of diffusion.

[73] Kelley & Harris (2020) supra nota 68. 1
[74] Chesney & Citron (2019) supra nota 22. 1783
[75] Venema, A. E., & Geradts, Z. J. (2020). Digital Forensics, Deepfakes, and the Legal Process. Scitech Lawyer, 16(4), 14-23. p. 16
[76] Chesney & Citron (2019) supra nota 22. 1785
[77] ibid
[78] Ryan, Patrick.(2020) "'Deepfake' Audio Evidence Used In UK Court To Discredit Dubai Dad". *The National*, https://www.thenationalnews.com/uae/courts/deepfake-audio-evidence-used-in-uk-court-to-discredit-dubai-dad-1.975764.

Increasing fidelity of deepfakes will no doubt damage the traditional credibility of audiovisual evidence in courts. Furthermore, if increased skepticism because of deepfake becomes prevalent enough, it would cause judges as well as other participants of the judiciary to lose their belief that the truth can be discovered. And It would be a devastating blow to the fundamental principles the judicial system built upon.

Although it is very little discussed in the literature, the harm that deepfakes could inflict on the judiciary is not just limited to the abovementioned two scenarios. There is a third possible scenario in which deepfakes could directly target judges and jurors, thus threatening the authority of the judiciary. Imagine a deepfake video released in the wild in the wake of the court proceedings regarding matters of public interests, depicting the judge taking a bribe to favor the defendant. Although the aforementioned two scenarios are more or less related to the audio-video evidence admissibility standards of courts, this third scenario concerns the authority and impartiality of the judiciary, which is a potential ground for interference with free expression.

2.3 Contribution of Social Media to the problem of Deepfake.

Platforms started banning or removing pornographic deepfakes when they first appeared and proliferated online.[79] Social media facilitated the spread of misinformation, as it lets the spreader circumvent "the conventional 'gatekeeping mechanisms, such as professional editors."[80]
As a matter of fact, we live in a "post-truth" era, where malevolent actors increasingly rely on digital disinformation campaigns in order to manipulate public opinion.[81] Social media platforms are perfect places for such disinformation campaigns as they provide instant delivery of content without verification.[82] These digital platforms completely changed the content creation dynamics by creating a landscape where anyone can easily produce, share and circulate the content instantly, whereas before the content creation and delivery to the intended audience required some journalistic knowledge and investment. Nowadays, social media platforms enable users to share

[79] Vincent, James (Feb. 7, 2018) Twitter is Removing Face-Swapped Al Porn from its Platform, too, VERGE

[80] Lewandowsky, S., Ecker, U., Seifert, C., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131. p.110

[81] Westerlund. (2019) supra nota 10. 39

[82] Nicolas, A. C. (2018). Taming the Trolls: The Need for an International Legal Framework to Regulate State Use of Disinformation on Social Media. Georgetown Law Journal Online, 107, 36-62. p.42

any kind of information they want and to reach any audience they want. This inevitably brought new dimensions to free expression rationales in the online environment.[83]
The algorithms these platforms built upon are aiming at maximizing user engagement at all costs. These algorithms relying on the existing human biases shape the landscape of social media platforms in a way that incentivizes provoking content.[84]

Also, human beings tend to accept and share the information supporting their preexisting beliefs. This is called "confirmation bias," [85] which refers to the psychological inclination to find and accept information that is aligned with persons' preexisting beliefs, opinions and to interpret that information according to these beliefs, opinions.[86] Algorithms heavily rely on that cognitive bias, thus creating "filter bubbles." These filter bubbles contribute to the circulation of false information, as it regularly fills the users' news feed with the information that approves their preexisting beliefs.

Furthermore, as a natural tendency, human beings are more inclined to spread negative and new information, and this can speed up deepfake circulation. The study conducted on Twitter found that hoaxes and false rumors spread ten times faster than true information.[87] According to this study, "the top 1% of false news cascades diffused to between 1000 and 100,000 people, whereas the truth rarely diffused to more than 1000 people."[88] It is because people tend to share information that is novel and that elicits more extreme reactions such as surprise and disgust.[89]

Combined with the new dimensions created by social media in news and content sharing, deepfakes have an immense capability to wreak havoc in the already shaken democratic values.

[83] Dias Oliva, T. (2020). Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. *Human Rights Law Review*, *20*(4), 607-640. https://doi.org/10.1093/hrlr/ngaa032
[84] Waldemarsson, C. (2020). Disinformation, Deepfakes & Democracy The European response to election interference in the digital age (pp. 6-7). The Alliance of Democracies Foundation.
[85] Waldman, A. (2018). The marketplace of fake news. University of Pennsylvania Journal of Constitutional Law, 20(4), 845-870. p.851
[86] Ling, R. (2020). Confirmation bias in the era of mobile news consumption: the social and psychological dimensions. *Digital Journalism*, *8*(5), 596-604. p.596
[87] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146-1151. p.1146
[88] Ibid.
[89] Ibid.

The combination of cognitive biases and algorithmic filter bubbles boosts the circulation of deepfakes. Furthermore, the borderless nature of the internet and storing data online makes it near impossible to erase the content once it is shared online.[90]

[90] Chesney & Citron (2019) supra nota 22. 1774

3. EUROPEAN PERSPECTIVE OF FREEDOM OF EXPRESSION

In the United States, academic literature examining legal responses to Deepfakes from the perspective of the First Amendment is abundant. But unfortunately, the same trend is not observed in Europe. The studies focusing on the Freedom of Expression and Deepfake regulation rationales are rare. Before signatory states rush to regulate the fake news and deep fake problem, it is of utmost importance to assess the situation from the perspective of ECHR, as well as the related case law of ECtHR.

Article 10 of the Convention reads as follows:

"1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.

2. The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary."

Article 10 does not limit the forms and means in which information and ideas could be created, disseminated, and received.[91] This means that all forms and means of expression, including the ones expressed through deepfakes, are protected under the convention. In the context of the right to freedom of expression, deepfakes should be considered as a form or a means of expression. This understanding explains why regulation of deepfakes could impair the right to free expression.

Before discussing the restrictions on the right to freedom of expression, a state's relation to this right should be taken into account. It is the state that will provide legal protection to free expression. Indeed, the state has an effective function in realizing the freedom of expression.

[91] McGoldrick, D. (2013). The Limits of Freedom of Expression on and Social Networking Sites: A UK Perspective. *Human Rights Law Review, 13*(1), 125-151. p.126

Moreover, within the framework of pluralist democratic principles, the state should even ensure the protection of the expression that is not aligned with the norms set by the state itself. In this context, besides making the necessary arrangements in the formation phase of the individual's thought, the state should not condemn the individuals for expressing and spreading their opinions and should provide individuals with an environment where they can act in accordance with their own opinions within legitimate limits. Therefore, it appears that the state has two types of obligations, one positive and the other negative. In accordance with its positive obligation, the state prepares the environment in which this freedom can be enjoyed. As a negative obligation, a state must not interfere with the exercise of this freedom within its limits.

As it can be seen, Art. 10 paragraph 2 of ECHR lays down the conditions in which freedom of expression might be limited. Nevertheless, these conditions and grounds for limitations are subject to very strict scrutiny of ECtHR. The court applies the 3-pronged test to assess the legitimacy of any intrusion:[92]

1) The interference must be prescribed by law (the rule of law, legal assurance against arbitrariness),

2)The aim should be the protection of at least one of the interests specified in art.10(2),

3) At the same time, it should be necessary in a democratic society.

The respondent state must prove that all three conditions are fulfilled. The ECtHR examines these three conditions in the aforementioned order. When it detects that one of the conditions has not been fulfilled, it stops the examination of the file, decides that the interference in question was unfair and thus freedom of expression was violated.[93]

One of the important criteria foreseen is that the limitation cannot be general. Any restriction, condition, limitation, or any form of interference on freedom of expression can only be applied to a particular exercise of that freedom. The content of the right to freedom of expression can never be touched.[94] In this respect, Article 17 of the Convention states:

"Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention."

[92] Bychawska-Siniarska (2017) supra nota 9. 32

[93] Ibid.

[94]Ibid.

A limitation on the content of a right is like the destruction of that right. At the same time, the national authorities are not obliged to intervene on any of the grounds enumerated in paragraph 2, as this would mean a restriction on the content of the right in question. For example, damage to a person's reputation should not be seen as a crime or cause of compensation in all cases.[95]

Similarly, public statements that would endanger the authority of the judiciary should not be punished in every case. In other words, the public authorities have no obligation to determine and implement a restriction or punitive measures on the exercise of the right to freedom of expression. It can be described as a possibility rather than an obligation. Accepting it as an obligation would lead to a hierarchy of rights and values or interests and result in freedom of expression being placed at the bottom of the list, after, for example, the right to protect human dignity, morality, or public order. Such a hierarchy would also be against the Convention, which ensures equality of rights and does not allow permanent restrictions on the exercise of a right.[96]

3.1 Prescribed by law

The most basic condition of the restrictions set out in paragraph 2 of Article 10 of the European Convention on Human Rights is that restrictions and sanctions must be prescribed by law. Under this condition, any interference with the exercise of freedom of expression must have a basis in the laws of the country. The meaning of the condition that the intervention is prescribed by law in the case law of the ECtHR; means that the measure or act interfering with the right or freedom has a legal basis in national law. As a rule, this means that it is a written and public law adopted by the parliament. Whether such a restriction should be possible must be decided by the country's parliament.[97] Furthermore, the law containing interference with the freedom of expression should meet qualitative standards developed by the ECtHR jurisprudence.[98] These quality standards require the law to be "public, accessible, predictable and foreseeable."[99] This also means that the consequences that may arise as a result of a particular action can be reasonably predicted in the

[95] Ibid.

[96] Macovei, M. (2004). A guide to the implementation of Article 10 of the European Convention on Human Rights. Human rights handbooks, (2). p.36.

[97] Bychawska-Siniarska (2017) supra nota 9. 39

[98] Mendel, Toby. (2012) "freedom of expression: a guide to the interpretation and meaning of article 10 of the european convention on human rights." *council of europe* . p.34-35 cited in Todorova, Elena. (2020) "restrictions on freedom of expression-necessity in a democratic society." *vizione* 34. 69-80. p.72

[99] Bychawska-Siniarska (2017) supra nota 9. 40

circumstances of the event with the help of a consultant. Foreseeability does not have to be of an absolute degree. Legally foreseeability, although desirable, can lead to excessive rigidity. However, the law should be able to adapt itself to change. Many laws are interpretative formulas, the interpretation and application of which, by the nature of the work, depend on practical reality. Again, foreseeability depends to a large extent on the content of the text in question, the nature of the area it covers, and the number of people it covers. Legal foreseeability does not conflict with requiring the person to use the assistance of a consultant to obtain reasonable information about the consequences of a particular action. This rule also applies especially to those who have to take great care in their professional life; they may be expected to be more careful in assessing the risks they are likely to face.[100]

3.2 The Legitimate Aim

ECHR art. 10(2) lists the legitimate grounds on which free expression may be limited, and member states cannot interfere with this right on any other grounds.[101] ECHR stipulates that the restriction must have a legitimate aim. Paragraph 2 of Article 10 of the Convention enumerated legitimate aims. The obligation to show that the limitation of freedom of expression is justified by legitimate aims lies with the State. ECtHR stated that the interference with the freedom of expression is considered a violation if it does not fall under exceptions listed in art. 10(2).[102] The legitimate aim must be put forward convincingly, and there must be a "pressing social need." Useful, desired, acceptable, ordinary situations do not mean "necessary." Restricting freedom of expression should be proportionate to the legitimate aim pursued.

3.3 The necessity in a democratic society

Restricting the exercise of freedoms in articles 8, 9, 10, and 11 of the ECHR should correspond to the necessities of a democratic society. The European Court of Human Rights has expressed its general approach to the “necessity” test and the margin of appreciation in the Silver and Others v. England case. According to this decision:[103]

[100] The Sunday Times V. The United Kingdom (No. 1) No. 6538/74 ECHR 1979. Para 45
[101] Bychawska-Siniarska (2017) supra nota 9. 40
[102] Handyside V. The United Kingdom No. 5493/72 ECHR 1976. para 43
[103] Silver And Others V. The United Kingdom No. 5947/72 ECHR 1983 para 97

1) “the adjective "necessary" is not synonymous with "indispensable," neither has it the flexibility of such expressions as "admissible," "ordinary," "useful," "reasonable," or "desirable."

2) “the Contracting States enjoy a certain but not unlimited margin of appreciation in the matter of the imposition of restrictions, but it is for the court to give the final ruling on whether they are compatible with the Convention;.”

(c) “the phrase "necessary in a democratic society" means that to be compatible with the Convention, the interference must, inter alia, correspond to a "pressing social need" and be "proportionate to the legitimate aim pursued.;”

(d) “those paragraphs of Articles of the Convention which provide for an exception to a right guaranteed are to be narrowly interpreted.”

Freedom of expression is one of the fundamental foundations of a democratic society and constitutes one of the basic conditions for the progress of the democratic society and the self-development of each individual. Protection of Freedom of expression within the boundaries of Article 10 is not only applied for "information" and "ideas" that are deemed to be favorable or regarded as harmless or unworthy of attention, but also for "information" and "ideas" that offend, shock or disturb. These are the requirements of pluralism, tolerance, and open-mindedness; Without them, there will be no democratic society.[104] Freedom of expression is subject to some exceptions in article 10 of the Convention; however, these exceptions should be interpreted narrowly, and the necessity of restrictions should be convincingly demonstrated.

The concept of "necessary" in the sense of art. 10(2) of the Convention indicates the existence of "pressing social need." The Contracting States have a certain margin of appreciation in assessing whether such a need exists; national discretion goes hand in hand with European supervision. This audit includes both the legislation and the court decisions implementing this legislation, even if given by independent courts. The ECtHR, therefore, has the final say to decide on whether a "restriction" is compatible with the freedom of expression protected by Article 10 of the Convention.[105]

ECtHR, in The Sunday Times V. The United Kingdom case, stated that” the court's task, in exercising its supervisory jurisdiction, is not to take the place of the competent national authorities but rather to review under Article 10 (art. 10) the decisions they delivered pursuant to their power of

[104] Lingens V. Austria No. 9815/82 ECHR 1986. Para 41

[105] Ibid. para 39

appreciation. This does not mean that the supervision is limited to ascertaining whether the respondent State exercised its discretion reasonably, carefully, and in good faith; what the court has to do is to look at the interference complained of in the light of the case as a whole and determine whether it was "proportionate to the legitimate aim pursued" and whether the reasons adduced by the national authorities to justify it are "relevant and sufficient."[106]
In doing so, the court must be convinced that the national authorities are applying the standards in accordance with the principles contained in article 10 of the Convention and also that they are based on an acceptable assessment of the relevant facts.[107]

The "necessity" specified in paragraph 2 of Article 10 means "pressing social need." As mentioned earlier, the national bodies of the Contracting States have a certain margin of appreciation in assessing whether this need exists. However, the decisions and actions of all national bodies, including decisions made by an independent national court, are under the control of the European Court of Human Rights. For this reason, the European Court of Human Rights has the authority to make the final decision on whether a restriction is compatible with freedom of expression, which is guaranteed by Article 10 of the Convention.
In order to determine whether the interference with freedom of expression constitutes mandatory measures in a democratic society, national courts must first apply the principle of proportionality. The question to be answered here is: "Is the aim proportional to the means used to achieve that aim?" In this equation, the "aim" is one or more of the values or interests enumerated in paragraph 2 that allow the states to interfere with freedom of expression for the sake of their protection.[108]
The "means" is the intervention itself. Thus, "aim" is the specific interest put forward by the state, such as "national security," "public order," "morality," "the rights of others." "means" is a specific measure adopted or applied against an individual exercising his freedom of expression. For example, "means" can be one of the following: criminal conviction for insult and defamation, sentencing to pay compensation, searching a newspaper office, seizing the means used to express an opinion, content removal or blocking, and so on.[109]

The decision on proportionality is based on the principles governing a democratic society. In order to prove that interference is "necessary in a democratic society," the court must be convinced of

[106] The Sunday Times V. The United Kingdom (No. 2) No. 13166/87 ECHR 1990. Para. 50.
[107] jersild v. Denmark No 15890/89. ECHR 1994. Para 31.
[108] Bychawska-Siniarska (2017) supra nota 9. 44
[109] Macovei, M. (2004). supra nota 96. 66-67

the existence of a "pressing social need" that would entail that specific restriction on the exercise of freedom of expression. As in the case of "Observer and Guardian v. England," the ECtHR stated that the adjective "necessary" in the meaning of paragraph 2 of Article 10 involves the existence of a "pressing social need,"[110] the question should be asked here is whether threats of deepfakes creates “pressing social need” for intervention. But unfortunately, it is not possible to give a straight answer to this question because the case law of the court, although has some general standards of examination, is unique in every case. The ECtHR examines every particular case as a whole, meaning that the specific situation and circumstances surrounding the interference with free expression are taken into account. Therefore, the particular interference with the right could amount to a violation in one case, but in another could be considered necessary. But nevertheless, threats of deepfakes discussed in chapter II could create “pressing social need” on the grounds of national security, territorial integrity, public order, reputation and rights of others, and authority of the judiciary.

[110] Ibid 67

4. DETERMINING THE GROUNDS FOR INTERFERENCE

4.1 National Security and Territorial integrity

The ECtHR states that it is possible to limit freedom of expression as a method of protecting national security, maintaining public order, and combating terrorism. However, it emphasizes that the limitation to be made based on these reasons can be accepted if there is a proportionality between the means used for the aim to be obtained and there is a pressing social need exists in this direction. In Zana v. Turkey case, the court stated that if the member states of the convention interfere with the freedom of expression by using this discretion, the delicate balance between the freedom of expression of individuals and the right of states to protect themselves from the actions of terrorist organizations will also need to be struck.[111]

The case concerned the interview of Mr. Zana published in the daily newspaper "Cumhuriyet" on 30 August 1987. In his statement, Mr. Zana supported the terrorist organization PKK, which conducted violence and massacres in southeast Turkey. In his statements, Mr. Zana expressed his support for the "PKK national liberation movement" while at the same time said that he is not "in favor of massacres." Furthermore, he also said that "Anyone can make mistakes, and the PKK kill women and children by mistake."[112]

The court found Mr. Zana's statements contradictory and ambiguous stating that, "it would seem difficult simultaneously to support the PKK, a terrorist organization which resorts to violence to achieve its ends, and to declare oneself opposed to massacres; they are ambiguous because whilst Mr. Zana disapproves of the massacres of women and children, he at the same time describes them as "mistakes" that anybody could make."[113]

Court also stated that, when deciding whether there was a "necessity" for interference, the case at hand should be examined as a whole. Therefore, the statement of Mr. Zana cannot be assessed in isolation. The raging situation in southeast Turkey, where PKK conducted violence and massacres,

[111] Zana v. Turkey no 18954/91 ECHR 1996. para. 55
[112] Ibid. para. 57
[113] Ibid. para. 58

should be taken into account. In these extreme circumstances, an interview given by Mr. Zana would have exacerbated the tensions in the region even more. The court acknowledged that the interference(imprisonment of Mr. Zana) was answering the "pressing social need."[114]

Zana v. Turkey case is of great importance in this research, as it helps to assume that in a similar situation, where deepfake video could further increase the tensions and incite violence, the interference(even imprisonment as a criminal sanction) would be considered "necessary in a democratic society."

In addition, in Jersild v. Denmark case, the court reiterated that "… it is commonly acknowledged that the audiovisual media have often a much more immediate and powerful effect than the print media... The audiovisual media have means of conveying through images meanings which the print media are not able to impart.".[115] This statement further cements the assumption put forward above.

Another relevant case law is the Surek v. Turkey case[116]. In this case, the applicant was convicted of disseminating separatist propaganda threatening national security, territorial integrity, and public order. The applicant's journal had published two letters submitted by readers. These letters strongly condemned the military activities of the authorities in southeast Turkey and accused them of persecuting the Kurdish people in their struggle for freedom and independence.

The first letter, titled "Weapons Cannot win against Freedom," refers to the two alleged massacres deliberately conducted by the authorities as part of the strategic campaign to eradicate the Kurds. It ends by repeating the determination of the Kurds to gain their freedom. The second letter, titled "It is our fault," alleges that the institutions of the Republic of Turkey turned a blind eye to imprisonment, torture, and killing of dissidents in the name of the protection of democracy and the Republic.[117]

The Istanbul National Security Court found that the charge against the applicant under Article 8 of the Anti-Terror Law No. 3713 of 1991 was proven. The court noted that the letters threatened the territorial integrity of the Republic of Turkey by describing the southeast regions as an independent "Kurdish" state and the PKK as a liberation movement.[118]

[114] Ibid. para. 59-62

[115] Jersild v. Denmark. Supra nota 107. para.31

[116] Sürek V. Turkey (No. 1) No 26682/95 ECHR 1999.

[117] Ibid para 60

[118] ibid

The ECtHR pointed out that there is a very narrow scope in Article 10(2) of the Convention for limiting political speeches or discussions on public interest issues. Also, the limits of permissible criticism are wider with regard to the government than with respect to private citizens or politicians. "Movements in a democratic system or omissions of the government should be under close scrutiny not only of the legislative and judicial authorities, but also of the public." In addition, the dominant position of the Government requires restraint on the use of criminal proceedings, especially in cases where unjust attacks and other means of responding to criticism of their adversaries exist. But the ECtHR also acknowledged that, as guarantors of public order, states might adopt measures, even of a criminal-law nature, intended to react appropriately and without excess to such remarks. And authorities have a wider margin of appreciation when such remarks incite violence against an individual, public official, or sector of the population.[119]

Again, as in the "Zana v. Turkey," ECtHR assessed the case as a whole and took into account the security circumstances and disturbances in the southeast of Turkey at the time the letters were published. The ECtHR thus found that, given the circumstances, the letters were capable of inciting violence "by instilling a deep-seated and irrational hatred against those depicted as responsible for the alleged atrocities. Indeed, the message which is communicated to the reader is that recourse to violence is a necessary and justified measure of self-defense in the face of the aggressor."[120]

Furthermore, The ECtHR stated that, although the fact that the "information" and "ideas" may offend, shock, or disturb does not justify the interference. Because, In the case at hand, the issue was "hate speech" and "glorification of violence."[121] Here, the court's approach to hate speech should be discussed as it lacks a uniform position. ECtHR considers hate speech unworthy of protection and, in doing so, employs different approaches of international instruments and national jurisdictions. But unfortunately, the court does not clearly define what is considered hate speech and what standards should be employed to identify it or assess the hate speech regulations.[122] The ECtHR's case law does not answer what is "hate based on intolerance" or what distinguishes "offensive speech" from "hate speech." When examining the hate speech laws, ECtHR refers to a

[119] Ibid para 61
[120] Ibid para 62
[121] Ibid para 58, 62
[122] Sottiaux, S., & Rummens, S. (2012). Concentric democracy: Resolving the incoherence in the European Court of Human Rights' case law on freedom of expression and freedom of association. *International Journal Of Constitutional Law*, *10*(1), 106-126. https://doi.org/10.1093/icon/mor074 p.111

wide variety of threats and harms that hate speech poses. But in doing so, ECtHR does not explain how these harms affect the proportionality test. For instance, the court stated that hate speech, as well as causing discrimination and violence, can threaten "social peace" and political stability" or instill hostile attitudes among people.[123] As can be seen, ECtHR's threshold for hate speech is very low, and it results in very "broad and vague" notion of hate speech. It is concerning because the low threshold can pave the way for legitimate free expression to be prosecuted under the stamp of hate speech. [124]

What is more interesting in this case law is that the ECtHR stated that the fact that the applicant is not the author of the letters and he only published them in the journal he owned without having an editorial relationship with them does not exempt him from liability. The reason is that he owned the journal and had the power to shape the editorial direction, and hence "duties and responsibilities" within the context of art.10(2).[125] The "duties and responsibilities" notion defined in Article 10(2) is unique, since it is not prescribed in any other Article of the Convention, even in those that have prescribed interference. The notion obliges the person enjoying the right to free expression to act responsibly and not to use it to abuse other rights enshrined in the ECHR.[126]

When deciding the liability of intermediaries, the ECtHR attributes a relatively wide margin of appreciation to signatory countries. This gives permission to states to apply strict liability rules for the liability of intermediaries in the context of third-party acts. This approach makes different content moderation rules on these platforms, such as comment filtering, removing, and blocking, compatible with the convention.[127] This is particularly important, as in the case of deepfakes threatening national security and territorial integrity, it gives power to states to hold accountable not just the creator but also the digital platforms who play a crucial role in the dissemination of that material. But increasing the fidelity of deepfakes will inevitably be able to avoid content monitoring and regulation mechanisms of digital platforms. In this case, signatory states might put pressure on these platforms and force them to take more drastic measures. This could create a situation where digital platforms might resort to more strict censorship mechanisms, thus creating the risk of silencing legitimate free expression as collateral damage.

[123] Feret v. Belgium No. 15615/07 ECHR 2009 cited in Sottiaux & Rummens (2012) supra nota 116. 111

[124] Ibid.

[125] Sürek V. Turkey supra nota 116. para 63

[126] Flauss, Jean-Francois. "The European Court of Human Rights and the Freedom of Expression." Indiana Law Journal, vol. 84, no. 3, Summer 2009, pp. 809-850. HeinOnline. p. 810

[127] Husovec, M. (2014). ECtHR rules on liability of ISPs as a restriction of freedom of speech. Journal of Intellectual Property Law & Practice, 9(2), 108-109.

4.2 Public Order and Prevention of Crime

According to the ECtHR, a state, as a guarantor of public order, if it deems necessary, can take, within limits, measures (even punitive) against criticisms directed to itself.[128] The contracting state has a wider margin of appreciation in interfering with freedom of expression if the criticism incites violence against individuals, public officials, or a section of society.[129]

The court's stance in this respect implies that as long as the expression does not incite violence or other unlawful actions, a state cannot interfere with freedom of expression. However, if the deepfake video, for example, depicting police brutality leads to or could lead to violent actions against police officers or public officials in general, that would be considered as a "pressing social need" for interference with free expression.

In the case Saszmann v. Austria, as the applicant incited the members of the army to disobey and violate the military laws of Austria, his conviction(imprisonment) was found to be justified. The reason is that inciting others to disobey the laws amounts to "unconstitutional pressure aiming at the abolition of laws which had been passed in a constitutional manner. Such unconstitutional pressure could not be tolerated in a democratic society."[130] This case also supports the view that if deepfakes are used to destruct public order and the rule of law, they are excluded from the protective sphere of article 10.

4.3 Protection of rights and reputation of others

Protecting the reputation and rights of others has been the "legitimate aim" used by national authorities more than any other reason to restrict freedom of expression. It has been used quite often to protect politicians and government officials from criticism. It is for this reason that the

[128] Incal v. Turkey no 22678/93 ECHR 1998. para.54
[129] Ceylan v. Turkey no. 23556/94 ECHR 1999. Para 34
[130] Saszmann v. Austria no 23697/94 ECHR 1997 cited in Dominika supra nota 39. 57

ECtHR has developed wide-ranging case law in this area providing high-level protection, in particular to the press. The privileged position of the press is due to the ECtHR's belief in the central role played by the expression of political opinions in a democratic society both in terms of the electoral process and everyday issues of the public interest. With regards to language, besides harsh and sharp criticism, the ECtHR considers colorful expressions that offer advantages in terms of drawing attention to the issues under discussion as acceptable.[131]

On the other hand, even if determined by positive law, objectively vulgar, obscene, offensive, insulting, degrading words and writings, swearing, libel, slander, and other expressions of thoughts aiming to hurt the reputation of others fall outside the scope of the law.

The prohibition of derogatory insults, cursing, slander, libel, and similar expressions against honor, dignity, the reputation of others, as they cannot be elements of free expression, should be regarded as the purification of free expression from foreign elements. It should also be remembered that such statements are no less bad than physical assault and should not be allowed in civilized societies.

Constantinescu v. Romania case[132] is one of the many cases ECtHR delivered judgment on regarding the interference with the right to free expression on this ground. The case concerned Mr. Constanticescu's conviction of libel. Mr. Constantinescu, the president of the teachers' union, in his interview to a newspaper, while expressing his dissatisfaction with the speed of criminal investigations regarding three teachers of the union called them "delapidatori"(persons found guilty of fraudulent conversion).[133]

The court emphasized that even if the impugned statements are spoken in the framework of a debate about the independence of trade unions and the functioning of the judicial organization, which constitute matters of public interest, there are limits to free expression. Along with the special role played by him as a union representative, the applicant should have reacted within limits set and without defamatory statements, in particular, to protect the rights and reputation of others. Therefore, the court found no violation of article 10 and concluded that there was a "pressing social need" for interference.[134]

[131] Bychawska-Siniarska (2017) supra nota 9. 63

[132] Constantinescu v. Romania No 28871/95 ECHR 2000

[133] Ibid. para 13

[134] Ibid. para 72-78

In the light of the principles of the ECtHR, all laws that aim to protect politicians or senior officials in general, through special or more severe punishments, against insults or attacks on their reputation, particularly those from the press, are incompatible with Article 10. Where such provisions exist and are attempted by politicians, national courts should refrain from applying them.[135] Instead, general legal provisions on defamation and reputation may be applied. Furthermore, when the honor and reputation of politicians come into conflict with freedom of the press, the national courts must carefully apply the principle of proportionality and decide, based on the guiding principles provided by the ECtHR in cases such as Lingens v. Austria, whether the journalist's conviction is necessary in a democratic society.[136]

Furthermore, where domestic law requires the defense of "proof of truth" in cases of defamatory statements, national courts should refrain from seeking such evidence, taking into account the ECtHR's distinction between facts and value judgments. The court reiterated in its case law that "while the existence of facts can be demonstrated; the truth of value judgments is not susceptible of proof."[137] Moreover, the defense of good faith should be accepted in cases of defamation, which are mainly based on facts. If the author has sufficient reason to believe that certain information is correct at the time it is published, then it should not be punished.[138] In other words, whether a statement is a value judgment or a fact is of utmost importance in determining the degree of protection granted to a particular expression. A value judgment enjoys almost absolute protection so long as it has some factual basis and is expressed in good faith.[139]

Furthermore, when discussing "proof of truth" and facts, it should not be forgotten that, particularly in the press, unfounded rumours do not fall under the umbrella of free expression. For a rumor to be protected by Art. 10, it should have at least some factual basis, and the journalist should at least attempt to establish this factual basis.

The court in case Timpul Info-Magazin and Anghel v. Moldova stated that "as part of their role of a "public watchdog," the media's reporting on "'stories' or 'rumours' … or 'public opinion'" is to be protected where they are not completely without foundation. The lack of any detailed information … despite the applicant newspaper's attempts to obtain such details, and the other

[135] Bychawska-Siniarska (2017) supra nota 9. 65
[136] Ibid.
[137] Pfeifer v. Austria No. 12556/03 ECHR 2007. Para 46
[138] Lingens v Austria supra nota 104. Para 45,46
[139] Flauss (2019) supra nota 126. 817

uncontested facts raising legitimate doubts … could reasonably have prompted the journalist to report on anything that was available, including unconfirmed rumours".[140]

The court's stance in Constantinescu v. Romania and Lingens v. Austria provides that contours of acceptable criticism against politicians are wider than ordinary people. The Court in Lingens v. Austria case stated that:

"The limits of acceptable criticism are accordingly wider as regards a politician as such than as regards a private individual. Unlike the latter, the former inevitably and knowingly lays himself open to close scrutiny of his every word and deed by both journalists and the public at large, and he must consequently display a greater degree of tolerance."[141]

Nevertheless, politicians' obligation to accept wider criticism does not mean that they cannot defend themselves against untrue or misleading attacks directed to their reputation or exercise of power.[142]

In light of the abovementioned cases, imagine a deepfake video posted online depicting the politicians 'confessing' his "immoral" and "unethical" internal and foreign policy. I believe in this case, the creator and publisher of the video will have the protection of article 10, as they can claim in good faith that the video was created to draw attention to the subject that is in the public interest. But similar doctored videos depicting ordinary people doing or saying things that are likely to harm their reputation will not enjoy the same protection as deepfakes targeting politicians.

But before drawing a conclusion, deepfakes targeting the reputation of others should be examined closely. These deepfakes could be divided into two categories:

First category deepfakes are the ones that target the reputation of the person depicted therein. In these doctored videos, the person does or says things as if they are 'exposing' themselves. For example, a politician is doing or saying things as if he is not aware of being recorded. Real evidence exists in the case of Jim Acosta, a journalist of CNN who was depicted in a doctored video in which he was acting aggressively against a White House intern.[143] Another example is the video of U.S. House of Representatives Speaker Nancy Pelosi, in which she was depicted as if

[140] Timpul Info-Magazin and Anghel v. Moldova No. 42864/05 ECHR 2007. Para 36

[141] Lingens v Austria supra nota 104. Para 42

[142] Sanocki v. Poland No. 28949/03. ECHR 2007. para. 61

[143] Aratani, Lauren. (2018) "Altered Video Of CNN Reporter Jim Acosta Heralds A Future Filled With 'Deep Fakes'". *Forbes*, https://www.forbes.com/sites/laurenaratani/2018/11/08/altered-video-of-cnn-reporter-jim-acosta-heralds-a-future-filled-with-deep-fakes/?sh=2ce42bff3f6c. Accessed 7 Apr 2021.

she was drunk and slurring words.[144] The video was used by her opponents to prove that She was unfit for her position as a House Speaker.

The second category defamatory deepfakes are the ones that depict a person uttering derogatory statements toward another person. These kinds of deepfakes can be considered as libel right away and prosecuted accordingly. For example, an Italian satirical TV show produced a deepfake video depicting Mr. Matteo Renzi, Italy's prime minister, uttering insults against fellow politicians. After it was shared and spread online, the video caused outrage among the public, as people believed it was real.[145]

Although case law gives a way to assume that deepfake targeting politicians with harsh criticism to the point where it would be considered defamation if it targeted an ordinary person will be accepted within free expression, the circumstances of the particular case should also be considered. As ECtHR, in its all judgments discussed so far, reiterated that when determining the existence of "pressing social need," the particular case should be considered as a whole, taking into account the circumstances the interference occurred. Hence, such deepfakes disseminated online shortly before an election could cause tremendous harm to the reputation of the politician. And in this case, the politician will not be able to debunk it in such a short period of time. Such well-timed deepfakes could manipulate the election by changing the public opinion about that particular politician. In addition, elections are the most fragile moments of democracies. As such, under these circumstances, deepfakes targeting the reputation of politicians could amount to "pressing social need" for interference.

The court's stance regarding the internet in cases Węgrzynowski and Smolczewski v. Poland and Editorial Board of Pravoye Delo and Shtekel v. Ukraine further support the abovementioned arguments. In these cases, the court stated that "the Internet is an information and communication tool particularly distinct from the printed media, especially as regards the capacity to store and transmit information. The electronic network, serving billions of users worldwide, is not and potentially will never be subject to the same regulations and control. The risk of harm posed by content and communications on the Internet to the exercise and enjoyment of human rights and

[144] "Fact Check: "Drunk" Nancy Pelosi Video Is Manipulated". *Reuters*, 2020, https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated-idUSKCN24Z2BI. Accessed 7 Apr 2021.

[145] Buo, Shadrack Awah. "The Emerging Threats of Deepfake Attacks and Countermeasures." *arXiv preprint arXiv:2012.07989* (2020). Pp 1-5 p 1

freedoms, particularly the right to respect for private life, is certainly higher than that posed by the press. Therefore, the policies governing the reproduction of material from the printed media and the Internet may differ. The latter undeniably have to be adjusted according to technology's specific features in order to secure the protection and promotion of the rights and freedoms concerned".[146]

4.4 Authority and Impartiality of judiciary

As discussed earlier, Deepfakes can be used to discredit the authority and impartiality of judiciary by directly targeting judges and jurists. It potentially allows restricting freedom of expression in the context of deepfakes on the ground of authority and impartiality of judiciary. In this respect, one of the key cases is De Haes and Gijsels v. Belgium case[147], which concerned five articles published by applicants criticizing the decision made by the judges of the court of appeal in a divorce proceeding. In that divorce case, the judges decided to give the custody of two children to their father, who was a famous notary and was previously accused of sexual abuse of the two children. However, before the divorce, the sexual abuse case against the father had been dropped without further indictment.[148]

In their five articles, applicants criticized the judges and alleged that the judges ruled on the case in favor of the notary because they had close political relations with him. Upon publications of these articles, the judge applied to the civil court, which found the journalists guilty of defamation and thus doubting the authority and impartiality of the judiciary. The court sentenced the two journalists to pay the civil damages and to publish the judgment in 6 newspapers.[149]

In De Haes and Gijsels v. Belgium case, ECtHR found the breach of Article 10 of the convention, since the journalist conducted a substantial amount of research before publishing the articles and they were not totally devoid of factual basis. In addition, the issue at hand was incest and the courts' approach to the issue. This made the issue a matter of public interest; therefore, the public

[146] Węgrzynowski and Smolczewski v. Poland No. 33846/07 ECHR 2007 para 58 and Editorial Board of Pravoye Delo and Shtekel v. Ukraine No. 33014/05 ECHR 2011 para 63
[147] De Haes And Gijsels V. Belgium No. 19983/92 ECHR 1997
[148] Ibid. para 7
[149] Ibid.

had the right to be informed about it. In light of these facts, ECtHR ruled that the interference was not "necessary in a democratic society" and journalists' right to free expression was violated.[150]

But the most important thing, in this case, is that ECtHR acknowledged that the judiciary and its members must have public trust. It is, therefore, of utmost importance to protect them from allegations and attacks devoid of any factual basis. Furthermore, as opposed to politicians, members of the judiciary are not able to publicly respond to allegations or attacks directed to them because of their duty of discretion. In addition, the decision of ECtHR judges on the breach of Article 10 was not unanimous. Judge Matscher and judge Morenilla presented their partially dissenting opinions pointing out that there was not a breach of article 10 and the interference with the journalists' free expression was "necessary" and "proportionate." Judge Matscher backed his opinion stating:

"What I find fault within the press articles that gave rise to the decision imposing a penalty on the applicants - albeit a nominal one - is the insinuation that the judges who gave that decision had deliberately acted in bad faith because of their political or ideological sympathies and thus breached their duty of independence and impartiality, all with the aim of protecting someone whose political ideas appeared to be similar to those of the judges concerned. Nothing justified such an insinuation, even if it had been possible to discover the impugned judges' political opinions."[151]

Another case to be considered in respect to interference with free expression on the ground of impartiality and authority of the judiciary is Prager and Oberschlick v. Austria[152]. Judge Morenilla also repeatedly referred to this case to back his dissenting opinion.[153] In this case, ECtHR specifically emphasized the role of the judicial system in a democratic society and stated that the right balance should be stroke between "the role of the press in imparting information on matters of public interests" (one of which is the functioning of the judicial system) and the protection of the rights of others and "the special role of the judiciary."[154] ECtHR stressed that "as the guarantor of justice, a fundamental value in a law-governed State, it must enjoy public confidence if it is to be successful in carrying out its duties." The court, therefore, considered it crucial that to be able to effectively function, the judiciary must enjoy public confidence and to maintain this confidence

[150] Ibid. para 39-49
[151] Ibid. Dissenting opinion of judge Matscher
[152] Prager And Oberschlick V. Austria No. 15974/90 ECHR 1995
[153] De Haes And Gijsels V. Belgium supra nota 147. Dissenting opinion of judge Morenilla
[154] Prager And Oberschlick V. Austria supra nota 152. Para 34

it should be protected from unfounded attacks.[155] ECtHR also acknowledged that national courts have a certain margin of appreciation determining the necessity of interference for the sake of impartiality and authority of the judiciary. However, ECtHR specifically stated that when assessing the facts and determining the necessity of interference, national courts must take into account that "the press is one of the means by which politicians and public opinion can verify that judges are discharging their heavy responsibilities in a manner that is in conformity with the aim which is the basis of the task entrusted to them."[156]

ECtHR's specific emphasis on the "public trust in courts" in the above-discussed cases gives enough grounds to assume that expressions in the form of deepfake targeting directly members of judiciary could be subject to intervention on the ground on impartiality and authority of the judiciary. Another reason to believe this assumption is that deepfaked video of judges depicting them taking bribes, using racial slurs, or committing any other wrongdoing would inflict incomparably profound damage to public trust in the judiciary because, as opposed to printed media, the speed and the scale of spread on social networks combined with the human cognitive biases maximize negative effects this technology has. Besides, the specific status of the members of the judiciary in public makes them more susceptible to such attacks, as they are considered the representation of justice and simply do not have the luxury to commit even ethical wrongdoings. The Members of judiciary are the ones who should represent the perfect persona that as well as obeying and maintaining legal order, is able to identify the truth and sentence the ones who break the law. Politicians, in contrast, are in a more grey area, as they are participants of heated political debates and very often are targeted with statements that are defamatory in nature. Politicians, since they are constantly in the public arena, are able to respond to such defamatory statements, sometimes by using even more defaming statements.

Judges do not have equal opportunity to respond to such attacks proportionately, and that is why they should be afforded more protection in this respect. To that end, states probably would have a wider margin of appreciation in interfering with free expression in the context of deepfakes directed to members of judiciary.

[155] Ibid.

[156] Ibid.

5. PASSING THE “PROPORTIONALITY” TEST: METHODS OF INTERFERENCE

As discussed earlier, the “necessary in a democratic society” test has two parts. Firstly, there should be “pressing social need” for interference with free expression. Secondly, the interference itself should be proportional to this need. In other words, the “proportionality” test requires from state to choose the least intrusive method of intervention to achieve the legitimate aim pursued. To decide whether the interference is proportionate to the legitimate aim pursued, ECtHR considers the following factors: “the form of expression, the means of dissemination, risk of abuse, the nature, and severity of the sanctions imposed.”[157]

The problem of the chilling effect of interference on the free flow of information that is often discussed in the context of regulation of fake news is valid for deepfakes as well. The application of intrusive measures inevitably comes with the social cost because of its chilling effect on the free flow of information. The degree to which the public interest is perceived to outweigh the social benefits of free flow of information in eliminating the social damage fake news and deepfakes cause will vary depending on how much social costs different societies are willing to endure.[158]

To date, the number of different countries imposing legal restrictions on deepfakes is growing. On the other side of the Atlantic, Virginia state amended its harassment laws to cover sexual abuse cases involving deepfakes.[159] The state of California took another path and outlawed the creation and dissemination of doctored audio, video, and images of politicians within 60 days of an election.[160] In January 2020, China announced new rules for online audio and video content creation, making it a criminal offense to produce and disseminate fake news through artificial

[157] Helm, R., & Nasu, H. (2021). Regulatory Responses to ‘Fake News’ and Freedom of Expression: Normative and Empirical Evaluation. *Human Rights Law Review*, *21*(2), 302-328. https://doi.org/10.1093/hrlr/ngaa060 p. 313

[158] Ibid.

[159] *Virginia bans 'deepfakes' and 'deepnudes' pornography*. BBC News. (2019). Retrieved 5 May 2021, from https://www.bbc.com/news/technology-48839758.

[160] Paul, K. (2019). *California makes ‘deepfake’ videos illegal, but law may be hard to enforce*. the Guardian. Retrieved 5 May 2021, from https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce.

intelligence without any marks explicitly showing that the content is synthetic.[161] According to the new rules, in case of failure to comply, not just the creator but also the digital platforms on which the content is disseminated are liable.[162] If tailored precisely and carefully, legal restrictions could be a powerful tool to negate the negative effects of deepfake technology. However, it should be noted that legal responses solely cannot solve the problem. Without technological solutions, such as deepfake detection tools, it would be near impossible to identify the fake content and its creator to impose legal sanctions in the first place. Besides, just relying on legal regulation would inevitably put more pressure on free expression, thus making it more susceptible to violation.

Furthermore, the regulatory response, as well as being least invasive, should be effective in fighting the threats posed by deepfakes. For example, studies show that although being the least intrusive method, the information correction method is not very effective against fake news.[163] Because of human cognitive biases such as confirmation bias, people tend to believe the information aligned with their pre-existing beliefs and reject the information contradicting these beliefs. Therefore, once a person chooses to believe certain content, correction to that content announcing its falsity might have a little effect in changing that person's belief. But it does not mean that this method should be dismissed; it should be rather used combined with other methods to maximize the efficacy. As mentioned earlier, Information correction does not directly interfere with fake content; but instead, it reveals the falsity of the information. This method is very likely to pass the "proportionality" test because it does not bar the free flow of information. This method is consistent with the "marketplace of idea" approach, which claims that individuals are capable of identifying truth when confronted with questionable information.[164] Twitter is one of the popular digital platforms employing this method. Twitter's information correction practice can be summarized in the following four steps:[165] a) Synthetic or manipulated tweets are identified through notice; b) before sharing the content, a user is warned regarding its manipulated or

[161] Woollacott, E. (2019). *China Bans Deepfakes In New Content Crackdown*. Forbes. Retrieved 5 May 2021, from https://www.forbes.com/sites/emmawoollacott/2019/11/30/china-bans-deepfakes-in-new-content-crackdown/?sh=29826acf3537.

[162] Statt, N. (2019). *China makes it a criminal offense to publish deepfakes or fake news without disclosure*. The Verge. Retrieved 5 May 2021, from https://www.theverge.com/2019/11/29/20988363/china-deepfakes-ban-internet-rules-fake-news-disclosure-virtual-reality.

[163] Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, *28*(11), 1531-1546.

[164] Helm & Nasu (2021) Supra nota 157. 315

[165] Harvey, D. (2019). Help us shape our approach to synthetic and manipulated media. Twitter. Retrieved from https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html cited in Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting Deepfakes: Media and Internet Giants' Converging and Diverging Strategies Against Hi-Tech Misinformation. *Media And Communication*, *9*(1), 291-300. https://doi.org/10.17645/mac.v9i1.3494 p.296

synthetic nature; c) A link to credible sources is added to the disputed content explaining why the content is false and misleading; d) Removal of the manipulated or synthetic content when it is potentially harmful or threatening to the safety of people.

Another method to tackle the deepfake problem is called “notice and takedown,” which refers to a mechanism where an individual, organization, or state authority can request from a digital platform (or from internet intermediary in general) to remove or block access to content infringing an individual’s right or law in general.[166] Although originally “notice and takedown” practice was designed to fight copyright infringements in the online environment, in the last couple of years, it is as well used to tackle fake news and disinformation. The main advantage of notice and takedown mechanisms is that they provide far quicker and effective relief than the judiciary. Nevertheless, despite being easily accessible, quick, and effective, this method has a very serious disadvantage which is its high potential to result in censorship and over-blocking, thus damaging free expression rights. Notice and takedown method puts on intermediaries a burden of identifying and deciding on the content's unlawful nature, which often involves competing rights and interests. This means that in that regard, intermediaries are replacing the courts, thus leading to the transfer of judicial power to the private sector.[167]

One of the many countries employing the notice and takedown method is Germany. In 2017, Germany enacted the Network Enforcement Act (*Netzwerkdurchsetzungsgesetz*), which requires social networks to remove "manifestly unlawful" content within 24 hours of notice.[168] Failing to comply could lead to penalties of up to 5 million euros. Under the notion of “manifestly unlawful” content, the law refers to 18 different provisions of the German criminal code such as defamation, insult, public incitement to crime, incitement to hatred, dissemination of propaganda material of unconstitutional organizations.[169] But there is not any guidance for social platforms to help them to determine whether complained content is “manifestly unlawful” or not. For example, by their nature, defamation cases are very complicated, and the defamatory weight of the expression and the protection afforded to it may change depending on in what form and to whom it was directed. Furthermore, the law raises concerns regarding censorship and over-blocking, which can cause a

[166] Kuczerawy, A. (2019). From ‘Notice and Take Down’to ‘Notice and Stay Down’: Risks and Safeguards for Freedom of Expression. *The Oxford Handbook of intermediary liability online*. Chapter V
[167] Ibid.
[168] McMillan, I. (2019). Enforcement through the network: The network enforcement act and article 10 of the European convention on human rights. Chicago Journal of International Law, 20(1), 252-290. p.261
[169] Ibid.

chilling effect on free expression. Digital platforms are more inclined to remove or block more content than necessary to avoid penalties, rather than prioritizing users' right to free expression. Over-blocking as a collateral effect of not precisely and carefully tailored legal measures inevitably interferes with lawful content, and therefore fails to pass the foreseeability requirement of ECtHR.[170] Lack of clear guidance and vague definition of fake news allows social platforms and national authorities to enjoy unfettered discretion, which can result in a risk of abuse, thus "curtailing freedom of expression arbitrarily and excessively."[171]

Another method to combat deepfakes is criminal sanctions. Currently, the States of Texas and Virginia criminalize malicious creation and dissemination of deepfakes. According to a new bill passed by the Texas senate amending the Election, code states that "A person commits an offense if the person, with intent to injure a candidate or influence the result of an election: (1) creates a deep fake video; and (2) causes the deep fake video to be published or distributed within 30 days of an election."[172]

Texas is the first state criminalizing political deepfakes. Contrary to Texas, the state of California imposes civil liability on malicious political deepfakes. But from the perspective of deterrence effect, criminal liability is more powerful than civil liability.

These criminal sanctions could be a very powerful regulatory solution to deepfakes as the threat of punishment can deter individuals from the creation and dissemination of malicious deepfakes. The deterrent effect of criminal sanctions can greatly reduce the malicious use of the technology. If a deepfake is not produced in the first place, the risk of exposure and manipulation of public opinion is eliminated. However, it should be noted that there is a lack of specific empirical studies supporting the claim that the threat of criminal sanctions discourages and eliminates the creation or dissemination of deepfakes or fake news.[173] Rather, its alleged effect is based on the crime prevention effect of the threat of punishment under the criminal liability. This exact crime prevention effect is the main reason why criminal sanctions are employed by Texas and Virginia laws.[174] The same argument can be put forward regarding the draft federal law of the US, the so-

[170] Vladimir Kharitonov V. Russia No. 10795/14 ECHR 2020. para 37–46

[171] La Rue, F. (2011). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. para 31.

[172] Texas Senate Bill 751

[173] Helm & Nasu (2021) Supra nota 157. 323

[174] Ibid.

called DEEPFAKES Accountability Act[175], which, if passed, will require deepfake creators to add digital watermarks or textual descriptions showing that the content is fabricated. The act states that:

“Whoever knowingly violates subsection (a) (produces and disseminates deepfake without digital watermark)—

“(i) with the intent to humiliate or otherwise harass the person falsely exhibited, provided the advanced technological false personation record contains sexual content of a visual nature and appears to feature such person engaging in such sexual acts or in a state of nudity;

“(ii) with the intent to cause violence or physical harm, incite armed or diplomatic conflict, or interfere in an official proceeding, including an election, provided the advanced technological false personation record did, in fact, pose a credible threat of instigating or advancing such;

“(iii) in the course of criminal conduct related to fraud, including securities fraud and wire fraud, false personation, or identity theft; or

“(iv) by a foreign power, or an agent thereof, with the intent of influencing a domestic public policy debate, interfering in a Federal, State, local, or territorial election, or engaging in other acts which such power may not lawfully undertake;

shall be fined under this title, imprisoned for not more than 5 years, or both.”[176]

The act actually is aimed at promoting increased transparency in the information ecosystem, and therefore is worth considering even if it is difficult to implement it against a technologically sophisticated person who is able to create malicious deepfakes with high fidelity.[177] Furthermore, the act could play a crucial role in establishing malicious intent because not including digital watermarks would indicate that the creator’s intent might not be parody or satire or simply an expression of his/her opinion in good faith. Although such criminal provisions will be very hard to impose due to common technical issues law enforcement faces in investigating cybercrimes, the mere existence of these provisions could help law enforcement to determine whether particular

[175] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019. USA

[176] Ibid. section 2. (f)

[177] Collins, A. (2019). Forged Authenticity: Governing Deepfake Risks. Lausanne: EPFL International Risk Governance Center. p.21

conduct is considered a crime or not, thus ending uncertainty surrounding different malicious use cases of this technology.

The criminalization of deepfakes envisages liability for those who create doctored content for disseminating online with the intention or possibility of causing public disturbances. ECtHR, when dealing with such criminal provisions, instead of dismissing them on account of uncertainty or ambiguity, will likely expect, under "proportionality" requirement, the criminal prosecution to protect legitimate interests by including different safeguards into the provisions to limit the extent to which it is used and to minimize the risk of abuse. And most needed safeguards, in this case, are requirements of malicious intent and the threshold of harm. Malicious intent and threshold of harm by forming *mens rea* and *actus reus* of deepfake crime will allow determining necessity and proportionality of punishing the person who creates and disseminates doctored audiovisual material with the intent of, for example, endangering national security or causing violence in public. ECtHR, in this matter, will expect the interference to be proportional to the level of threat that creates "pressing social need" for intervention on the grounds of legitimate interests listed in article 10.

In spite of the limitations, incorporation of criminal provisions dealing with malevolent uses of deepfakes provides for state authorities enough breathing ground to enforce preventive regulations by targeting attempts to create and disseminate doctored audiovisual materials to cause public disturbances. The main hassle here is the compatibility of criminal provisions with foreseeability and requirements ECHR. For example, in the context of fake news, the use of criminal provisions sanctioning the creation and dissemination of fake news has been deemed disproportionate in many countries, such as South Korea[178], due to the vagueness of definitions and the potential for arbitrary interference.[179] Hence the same possibility of disproportionality is also valid for criminal sanctions designed for deepfakes. Therefore, legislation should draft future criminal provisions with maximum precision that would allow a person to foresee the legal consequences of his/her action. Furthermore, provisions criminalizing malicious use of deep fakes should contain a wide range of sanctions such as corrective works, restriction of freedom, fines, deprivation of the right to hold certain posts or to engage in a certain activity, and imprisonment in more serious cases. A

[178] Park, A., & Youm, K. (2019). Fake news from legal perspective: The united states and south Korea compared. Southwestern Journal of International Law, 25(1), 100-119.

[179] Helm & Nasu (2021) supra nota 156. 23

wide range of sanctions will provide for national courts enough space to choose the most proportional sanction depending on the seriousness of each case.

However, it should be noted that criminal liability is not a silver bullet to the problem. Enforcement of criminal provisions will be a major difficulty, especially when the crime has extra-territorial nature. If foreign actors are involved in the commission of deepfake crimes, such as manipulation of an election, prosecution of those actors will be very troublesome. It brings us to the fact, which was repeatedly mentioned in this chapter, that legal regulation alone would not provide enough efficacy in negating threats posed by deepfakes. Legal solutions should be accompanied by technological solutions such as sophisticated deepfake detection tools to eliminate collateral damage to free expression. Furthermore, digital literacy and awareness of the public on the matter must be increased so that the public has the necessary fact-checking skills to identify whether particular content is doctored or not.

CONCLUSION

The emergence of the deepfake technology caused significant panic among the public, as it has the potential to further blur the line between truth and falsity, thus threatening the core of modern democracies, the informed decision making. While academics and lawmakers rush to solve the problem, it should be kept in mind that to date, there is no significant evidence showing the extent of the damage deepfakes could inflict on democratic societies. It is mostly scholars who construct or predict different scenarios of malicious deepfake use. Hence, it is very complicated to design regulation when the extent of possible negative consequences of malicious use, especially considering the fact that any regulation of this technology will have an impact on free expression. But existing literature on the topic, despite being sparse, gives convincing reasons to believe that deepfakes could pose serious threats to individuals, organizations, and the public in general. In this respect, the threats posed by Deep fakes could be divided into two categories:
1) Individual-level threats which include revenge porn, harassment, blackmail, identity theft, and deepfake powered social engineering attacks.
2) Public-level threats, which include dissemination of fake news, manipulation of elections, deepening socio-political polarization and disrupting democratic discourse, endangering national security, territorial integrity, public order, and effective functioning of judicial system.

As mentioned earlier close relationship between deepfakes and freedom of expression makes it of utmost importance to look at the issue from the perspective of ECHR art.10 and relevant case law of ECtHR. In cases related to freedom of expression, ECtHR applies its three-part test to assess the existence of a violation. According to this legal doctrine, for a state to be able to interfere with freedom of expression, the following conditions should be met:
a) the interference must be prescribed by law (the rule of law, legal assurance against arbitrariness).
b) The aim should be the protection of at least one of the interests specified in art.10(2)
c) The interference should be necessary in a democratic society.

Under "Necessary in a democratic society," ECtHR stated that "the adjective "necessary" is not synonymous with "indispensable," neither has it the flexibility of such expressions as "admissible," "ordinary," "useful," "reasonable" or "desirable." For imposing any restriction, condition or formality on freedom expression there should be "pressing social need" for intervention. In addition, the interference should be proportionate to the legitimate aim pursued. Unfortunately, there are not any ECtHR cases related to interference with free expression in the context of deepfakes. Therefore, this research examined article 10 related cases of ECtHR where "necessity" for interference were justified on different grounds. After that, interference in the context of deepfakes was compared to those cases in order to determine when it could satisfy the " necessary in a democratic society" test. The findings of this research support the hypothesis that deepfake threats could create "pressing social need" for intervention on the grounds of protection of national security, territorial integrity, public order, rights and reputation of others, and maintaining impartiality and authority of the judiciary.

But the mere existence of "pressing social need" is not enough to interfere with individuals' right to free expression. The interference itself should pass the "proportionality" test. In other words, it should be proportionate to the legitimate aim pursued. Therefore, methods to combat malicious deepfakes should be designed with enough precision and great care to avoid a chilling effect on free expression as collateral damage, which is a breach of proportionality. As the main playground of deepfakes is digital platforms, the most suggested interference methods are information correction, notice and takedown, and criminal sanctions.

The information correction method is the least invasive method, and therefore, chances are high for it to pass the "proportionality" test. But its effectiveness against deepfakes is questionable and existing studies show that it may not be the most effective way to combat disinformation. Contrary to information correction, the notice and takedown method is very intrusive and could be more effective to stop the spread of deepfakes and disinformation in general. But in light of censorship and over-blocking concerns regarding Germany's Network Enforcement Act, it can be noted that excessive use of this method might have a chilling effect on free expression as collateral damage. Furthermore, putting extra-legal pressure on digital platforms without any clear guidance will make them remove more content than necessary in order to avoid high penalties and force them to replace the judiciary in determining and deciding on the unlawfulness of particular content.

Incorporation of specific provisions into criminal codes dealing with malicious use of deepfakes is another suggested method that was employed by the state of Texas to criminalize political deepfakes aimed at manipulation of elections. Besides, the draft federal DEEPFAKES Accountability Act of US is currently discussed in Congress, and if passed, the act will criminalize the creation and dissemination of deepfakes without digital watermarks or descriptions showing that the content is fake. Criminalizing deepfakes could be a very powerful tool because of its deterrent effect. However, to avoid collateral damage to free expression, criminal provisions should be designed with maximum precision so that a person can foresee the consequences of his/her actions. Furthermore, for providing proportionality, criminal provisions should include a wide range of sanctions. This would provide for national courts more freedom to choose the most "proportionate" sanction depending on the nature of each case.

However, it should not be forgotten that the criminalization of deepfakes will not solve the problem entirely, since law enforcement has to deal with the same problems that every cybercrime poses, such as exterritorial nature, difficulty to obtain and maintain digital evidence to identify the perpetrator. To effectively combat deepfake threats, legal and technological solutions, as well as awareness-raising in public, should be used combined. Otherwise, legal solutions alone could put immense pressure on free expression, which is not acceptable, as it would mean sacrificing the core of our democracy to protect it. This is not the way to secure the future of democratic societies.

LIST OF REFERENCES

Books

1. Bychawska-Siniarska, D. (2017). Protecting the right to freedom of expression under the European convention on human rights: A handbook for legal practitioners. Council of Europe.
2. Kuczerawy, A. (2019). From 'Notice and Take Down'to 'Notice and Stay Down': Risks and Safeguards for Freedom of Expression. *The Oxford Handbook of intermediary liability online*.
3. Macovei, M. (2004). A guide to the implementation of Article 10 of the European Convention on Human Rights. Human rights handbooks, (2).

Articles

4. Baur C., Albarqouni S., Navab N. (2018) Generating Highly Realistic Images of Skin Lesions with GANs. In: Stoyanov D. et al. (eds) OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. CARE 2018, CLIP 2018, OR 2.0 2018, ISIC 2018. Lecture Notes in Computer Science, vol 11041. Springer, Cham. https://doi.org/10.1007/978-3-030-01201-4_28
5. Brown, N. I. (2020). Deepfakes and the Weaponization of Disinformation. Virginia Journal of Law & Technology, 23, 1-59.
6. Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, *28*(11), 1531-1546.
7. Chesney, B., & Citron, D. (2019). Deep fakes: looming challenge for privacy,democracy, and national security. California Law Review, 107(6), 1753-1820.
8. Citron, D. K., & Franks, M. A. (2014). Criminalizing revenge porn. *Wake Forest L. Rev.*, *49*, 345.
9. Dias Oliva, T. (2020). Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. *Human Rights Law Review*, *20*(4), 607-640. https://doi.org/10.1093/hrlr/ngaa032
10. Flauss, Jean-Francois. "The European Court of Human Rights and the Freedom of Expression." Indiana Law Journal, vol. 84, no. 3, Summer 2009, pp. 809-850. HeinOnline. p. 810
11. Frid-Adar, M., Diamant, I., Greenspan, H., Klang, E., Amitai, M., & Goldberger, J. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing, 321*, 321-331.
12. Hall, H. (2018). Deepfake videos: When seeing isn't believing. Catholic University Journal of Law and Technology, 27(1), 51-76.
13. Harris, D. (2018-2019). Deepfakes: False Pornography Is Here and the Law Cannot Protect You. Duke Law & Technology Review, 17, 99-128.
14. Hayward, P., & Rahn, A. (2015). Opening Pandora's Box: pleasure, consent and consequence in the production and circulation of celebrity sex videos. *Porn Studies*, *2*(1), 49-61. https://doi.org/10.1080/23268743.2014.984951

15. Helm, R., & Nasu, H. (2021). Regulatory Responses to 'Fake News' and Freedom of Expression: Normative and Empirical Evaluation. *Human Rights Law Review*, *21*(2), 302-328. https://doi.org/10.1093/hrlr/ngaa060
16. Husovec, M. (2014). ECtHR rules on liability of ISPs as a restriction of freedom of speech. Journal of Intellectual Property Law & Practice, 9(2), 108-109.
17. LaMonaca, J. P. (2020). break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes. American University Law Review, 69(6), 1945-1988.
18. Lewandowsky, S., Ecker, U., Seifert, C., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131.
19. Ling, R. (2020). Confirmation bias in the era of mobile news consumption: the social and psychological dimensions. *Digital Journalism*, *8*(5), 596-604.
20. Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, *7*(4), 415-423. https://doi.org/10.1080/23268743.2020.1757499
21. McGlynn, C., & Rackley, E. (2017). Image-Based Sexual Abuse. *Oxford Journal Of Legal Studies*, *37*(3), 534-561. https://doi.org/10.1093/ojls/gqw033
22. McGoldrick, D. (2013). The Limits of Freedom of Expression on and Social Networking Sites: A UK Perspective. *Human Rights Law Review, 13*(1), 125-151.
23. McMillan, I. (2019). Enforcement through the network: The network enforcement act and article 10 of the European convention on human rights. Chicago Journal of International Law, 20(1), 252-290.
24. Meškys, L., Liaudanskas, G., Kalpokienė, J., & Jurcys, P. (2020). Regulating deep fakes: legal and ethical considerations. *Journal of intellectual property law & practice. Oxford: Oxford university press, 2020, vol. 15, iss. 1.*
25. Westerlund. M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review, 9*(11), 40-53. p. 40
26. Mostert, F. (2020). 'Digital due process': A need for online justice. *Journal of Intellectual Property Law & Practice, 15*(5), 378-389.
27. Naruniec, J., Helminger, L., Schroers, C., & Weber, R. (2020). High-Resolution Neural Face Swapping for Visual Effects. *Computer Graphics Forum, 39*(4), 173-184.
28. Nicolas, A. C. (2018). Taming the Trolls: The Need for an International Legal Framework to Regulate State Use of Disinformation on Social Media. Georgetown Law Journal Online, 107, 36-62.
29. Öhman, C. (2019). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology, 22*(2), 133-140.
30. Park, A., & Youm, K. (2019). Fake news from legal perspective: The united states and south Korea compared. Southwestern Journal of International Law, 25(1), 100-119.
31. Perot, E., & Mostert, F. (2020). Fake it till you make it: an examination of the US and English approaches to persona protection as applied to deepfakes on social media. *Journal of Intellectual Property Law & Practice*, *15*(1), 32-39.
32. Pfefferkorn, R. (2020). "deepfakes" in the courtroom. Boston University Public Interest Law Journal, 29(2), 245-276.
33. Post, R. C. (2018). Data privacy and dignitary privacy: Google spain, the right to be forgotten, and the construction of the public sphere. *Duke Law Journal, 67*(5), 981-1072.
34. Schroeder, J. (2020). Free expression rationales and the problem of deepfakes within the e.u. and u.s. legal systems. Syracuse Law Review, 70(4), 1171-1204.
35. Shin, H. C., Tenenholtz, N. A., Rogers, J. K., Andriole, K. P., Michalski, M. G., Schwarz, C. L., . . . Gunter, J. (2018). Medical image synthesis for data augmentation and

anonymization using generative adversarial networks. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11037*, 1-11.

36. Sottiaux, S., & Rummens, S. (2012). Concentric democracy: Resolving the incoherence in the European Court of Human Rights' case law on freedom of expression and freedom of association. *International Journal Of Constitutional Law*, *10*(1), 106-126. https://doi.org/10.1093/icon/mor074
37. Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting Deepfakes: Media and Internet Giants' Converging and Diverging Strategies Against Hi-Tech Misinformation. *Media And Communication*, *9*(1), 291-300.
38. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146-1151.
39. Waldman, A. (2018). The marketplace of fake news. University of Pennsylvania Journal of Constitutional Law, 20(4), 845-870.

Normative Acts

40. Council of Europe, *European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14*, 4 November 1950, ETS 5, available at: https://www.refworld.org/docid/3ae6b3b04.html [accessed 12 April 2021]
41. Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019. USA
42. Texas Senate Bill 751

Court Cases

43. Ceylan v. Turkey no. 23556/94 ECHR 1999.
44. Constantinescu v. Romania No 28871/95 ECHR 2000
45. De Haes And Gijsels V. Belgium No. 19983/92 ECHR 1997
46. Editorial Board of Pravoye Delo and Shtekel v. Ukraine No. 33014/05 ECHR 2011
47. Feret v. Belgium No. 15615/07 ECHR 2009
48. Handyside V. The United Kingdom No. 5493/72 ECHR 1976.
49. Incal v. Turkey no 22678/93 ECHR 1998.
50. jersild v. Denmark No 15890/89. ECHR 1994.
51. Lingens V. Austria No. 9815/82 ECHR 1986.
52. Pfeifer v. Austria No. 12556/03 ECHR 2007.
53. Prager And Oberschlick V. Austria No. 15974/90 ECHR 1995
54. Saszmann v. Austria no 23697/94 ECHR 1997
55. Silver And Others V. The United Kingdom No. 5947/72 ECHR 1983
56. Sürek V. Turkey (No. 1) No 26682/95 ECHR 1999
57. The Sunday Times V. The United Kingdom (No. 1) No. 6538/74 ECHR 1979
58. The Sunday Times V. The United Kingdom (No. 2) No. 13166/87 ECHR 1990.
59. Timpul Info-Magazin and Anghel v. Moldova No. 42864/05 ECHR 2007.
60. Vladimir Kharitonov V. Russia No. 10795/14 ECHR 2020
61. Węgrzynowski and Smolczewski v. Poland No. 33846/07 ECHR 2007
62. Zana v. Turkey no 18954/91 ECHR 1996

Other Sources:

63. Aratani, L. (2020). *George Floyd killing: what sparked the protests – and what has been the response?*. the Guardian. Retrieved 11 May 2021, from https://www.theguardian.com/us-news/2020/may/29/george-floyd-killing-protests-police-brutality

64. Beaumont, Peter et al.(2021) "How A Mob Of Trump Supporters Stormed The Capitol – Visual Guide." *The Guardian*, https://www.theguardian.com/us-news/2021/jan/07/how-a-mob-of-trump-supporters-stormed-the-capitol-visual-guide. Accessed 26 Mar 2021.

65. Braunstein, E. (2018). At this Holocaust museum, you can speak with holograms of survivors. Timesofisrael.com. Retrieved 12 April 2021, from https://www.timesofisrael.com/at-this-holocaust-museum-you-can-speak-with-holograms-of-survivors/.

66. Caulfield, AJ. (2020) "The Truth About Recreating Paul Walker For Fast And The Furious - Exclusive." *Looper.Com*, https://www.looper.com/184468/the-truth-about-recreating-paul-walker-for-fast-and-the-furious/. Accessed 16 Mar 2021.

67. Chandler, S. (2020). Why Deepfakes Are A Net Positive For Humanity. Forbes. Retrieved 12 April 2021, from https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity/?sh=6d89f7cc2f84.

68. Chesney, R., & Citron, D. (2018). *Disinformation on Steroids: The Threat of Deep Fakes*. Council on Foreign Relations. Retrieved 11 April 2021, from https://www.cfr.org/report/deep-fake-disinformation-steroids.

69. Collins, A. (2019). Forged Authenticity: Governing Deepfake Risks. Lausanne: EPFL International Risk Governance Center.

70. Dalí Lives: Museum Brings Artist Back To Life With AI - Salvador Dalí Museum. *Thedali.Org*, 2019, https://thedali.org/press-room/dali-lives-museum-brings-artists-back-to-life-with-ai/.

71. Damiani, J. (2019, September 3). A Voice Deepfake Was Used To Scam A CEO Out Of $243,000. Forbes. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=34f4033c2241

72. David Beckham Travels to the Future to Announce the End of Malaria - IVCC. IVCC. (2021). Retrieved 12 April 2021, from https://www.ivcc.com/david-beckham-travels-to-the-future-to-announce-the-end-of-malaria

73. Freedom House. (2017). Manipulating Social Media to Undermine Democracy. Retrieved from Freedom on the Net 2017: https://freedomhouse.org/report/freedom-net/freedom-net-2017

74. Geraint, R. (2019). *Opinion: How the technology behind deepfakes can benefit all of society*. UCL News. Retrieved 12 April 2021, from https://www.ucl.ac.uk/news/2019/nov/opinion-how-technology-behind-deepfakes-can-benefit-all-society

75. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144. p. 139

76. Hawaii Worker Who Sent Missile Alert Was '100% Sure' Attack Was Real. *The Guardian*, 2018, https://www.theguardian.com/us-news/2018/feb/03/hawaii-worker-sent-missile-alert-100-percent-sure-attack-real.

77. Kelleher, S. (2021). *AI Has Resurrected Salvador Dali And Now He's Your Museum Tour Guide*. Forbes. Retrieved 12 April 2021, from

https://www.forbes.com/sites/suzannerowankelleher/2019/05/13/ai-has-resurrected-salvador-dali-and-now-hes-your-museum-tour-guide/?sh=76c050ad2d97.

78. Kim, B., & Ganapathi, V. (2019). Lumi`ereNet: Lecture Video Synthesis from Audio.

79. Kleinfeld, Rachel, and Aaron Sobel. *Eu.Usatoday.Com*, 2021, https://eu.usatoday.com/story/opinion/2020/07/23/political-polarization-dangerous-america-heres-how-fight-column/5477711002/. Accessed 21 Mar 2021.

80. La Rue, F. (2011). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.

81. Lomas, N. (2020, December 8). *Reface grabs $5.5M seed led by A16z to stoke its viral face-swap video app*. Techcrunch. https://techcrunch.com/2020/12/08/reface-grabs-5-5m-seed-led-by-a16z-to-stoke-its-viral-face-swap-video-app/

82. Meneses, J. P. (2021). Deepfakes and the 2020 US elections: what (did not) happen. arXiv preprint arXiv:2101.09092. 1-13, p 2

83. Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, *1*. p.1

84. Nicholson, Chris. "A Beginner's Guide To Generative Adversarial Networks (Gans)." *Pathmind*, https://wiki.pathmind.com/generative-adversarial-network-gan. Accessed 26 Mar 2021.

85. Panyatham, Paengsuda (2021) "Deepfake Technology In The Entertainment Industry: Potential Limitations And Protections — AMT Lab @ CMU." *AMT Lab @ CMU*, https://amt-lab.org/blog/2020/3/deepfake-technology-in-the-entertainment-industry-potential-limitations-and-protections.

86. Paul, K. (2019). *California makes 'deepfake' videos illegal, but law may be hard to enforce*. the Guardian. Retrieved 5 May 2021, from https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce.

87. Romano, Aja. (2019) "Deepfakes Are A Real Political Threat. For Now, Though, They'Re Mainly Used To Degrade Women.". *Vox*, https://www.vox.com/2019/10/7/20902215/deepfakes-usage-youtube-2019-deeptrace-research-report. Accessed 1 Mar 2021.

88. Ryan, Patrick.(2020) "'Deepfake' Audio Evidence Used In UK Court To Discredit Dubai Dad". *The National*, https://www.thenationalnews.com/uae/courts/deepfake-audio-evidence-used-in-uk-court-to-discredit-dubai-dad-1.975764.

89. Sayler, Kelley M., and Laurie A. Harris. (2020) *Deep Fakes And National Security*. Congressional Research Service, 1-3, https://crsreports.congress.gov/product/pdf/IF/IF11333.

90. Smith, Hannah, and Katherine, Mansted.(2020) *Weaponised Deep Fakes: National Security And Democracy*. ASPI International Cyber Policy Centre, 1-21.

91. Vincent, James (Feb. 7, 2018) Twitter is Removing Face-Swapped Al Porn from its Platform, too, VERGE

92. *Virginia bans 'deepfakes' and 'deepnudes' pornography*. BBC News. (2019). Retrieved 5 May 2021, from https://www.bbc.com/news/technology-48839758.

93. Waldemarsson, C. (2020). Disinformation, Deepfakes & Democracy The European response to election interference in the digital age (pp. 6-7). The Alliance of Democracies Foundation.

APPENDICES

Appendix 1. Non-exclusive licence

A non-exclusive licence for reproduction and publication of a graduation thesis[1180]

I _____Ibrahim Mammadzada___ (*author's name*)

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis

____________Deepfakes And Freedom of Expression: European Perspective___________,
(*title of the graduation thesis*)

supervised by_______________________Agnes Kasper___________________
(*supervisor's name*)

1.1 to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

_______________ (date)

[180] *The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.*