Legal Opinion

DeepFakes and the Collapse of Truth: a comparative analysis between American and European reactions to a new systemic threat

"One believes things because one has been conditioned to believe them."

— Aldous Huxley, Brave New World

Abstract

Nowadays, deepfakes allow anyone to distort reality by creating highly realistic images or videos depicting a person doing or saying things that never actually happened. Thus, this disturbing technology can be defined as a new type of "digital forgery" with multiple negative consequences for both the individual and society. The potential of deepfakes to disseminate completely false content and threaten an already vulnerable information digital ecosystem is immense, and deeply alarming. In light of the negative and positive aspects of this technology, the paper aims to provide a detailed analysis of the current legal framework in the US and Europe, taking into account and discussing potential remedies to address deepfakes. It will also discuss the crucial relationship between freedom of expression and the regulation of deepfakes, arguing for a revision of the traditional legal arguments underlying the regulation of online information flows.

Table of Content

Introduction

The age we currently live in has been defined the "post-truth era", in which deception has become the modern way of life and truth is under attack. The reality is not that far from a sinister and disturbing dystopian world, where anyone can distort what is real. This manipulative capacity has taken an exponential leap with "deepfakes", an artificial intelligence-based technology used to generate and alter digital content in a way hardly perceptible to the human eye. Thus, the falsehood dramatically infects and blurs with reality. As a result, it is extremely difficult to discern the actual truth from falsehood and we can no longer rely on vision/perception to form our beliefs.

Through innovative deep-learning techniques, deepfakes allow us to represent certain individuals in situations and attitudes that have never occurred in reality. Accordingly, the misuse of technology is undeniably powerful: deepfakes can be deliberately used as real cyber-weapons to cause a wide range of damage to both the individual, and society as a whole. This is even more alarming considering the fact that the tools that create deepfakes are easily usable and widely available to anyone. Moreover, such content is generally spread online and consequently become immediately ubiquitous on various social platforms. The harmful impact can reach systemic proportions. Consequently, the need to effectively act with an appropriate defence strategy is imperative.

The purpose of the dissertation is therefore to provide an analysis of the legal implications of this new technology and to discuss possible legislative solutions, with particular regard to the American and European systems.

First of all, the technical aspect of deepfakes and the types of damage that they can spread from them, as well as their possible positive applications, will be assessed. This first chapter will prove that in some cases the use of deepfakes is defended by freedom of expression guarantees, and therefore an absolute prohibition is not desirable as it would undermine and inhibit the potential benefits of this technology. In Chapter II, the different approaches to freedom of expression in the US and the EU will be reviewed, arguing for a possible reform of traditional legal arguments in the light not only of deepfakes, but also of the general dynamics of the online information market. Then, the third part of the dissertation discusses the measures currently in place to address the spread of deepfakes, and examines the legal response of the United States and the European Union, highlighting the latter's backwardness on this issue. Finally, a brief conclusion with some recommendations will follow.

CHAPTER I

The Growing Problem of Deepfakes

In order to effectively analyze the most appropriate legal remedies to approach deepfakes, it is necessary to understand what a deepfake is, how this technology works and what damage it can cause. Moreover, for a complete understanding of the issue, the potential positive applications of deepfakes should also be considered.

1. Deepfakes at a glance

The term deepfakes, coined from the combination of "deep-learning" and "fake", refers to digitally manipulated videos, images and sounds made with a sophisticated form of artificial intelligence to depict a certain person while performing actions that never actually happened.[1] Such synthetic content can be generated using a variety of deep-learning techniques and this method is what differentiates deepfakes from other false videos, which are 'manually' created. Notably, deep learning is a subset of machine learning that uses algorithms capable of simulating the functioning of the human brain in data processing and decision making. These algorithms develop a neural network, which learns from unstructured or unlabelled data sets.[2] Currently, the main technique for the creation of deepfakes is the Generative Adversarial Network (GAN), due to its flexible applications and realistic outputs.[3] The quality and sophistication of GAN is improving to such an extent that the work produced will be almost indistinguishable from the authentic video and it is likely to be able to circumvent any detection.[4] The key idea of GANs is to train two neural networks, generator and discriminator, together, in an adversarial relationship.[5] The generator analyzes a multitude of data samples from a particular source and creates a deepfake that tries to trick the discriminator into believing it is real.[6] On the other hand, the discriminator attempts to evaluate the quality of the clip and to detect the synthetically generated content. In this way, these two networks interact until the

[1] See Grace Shao 'What 'deepfakes' are and how they may be dangerous' (2019) CNBC https://www.cnbc.com/2019/10/14/what-is-deepfake-and-how-it-might-be-dangerous.html accessed 15 June 2020.
[2] See Alan Zucconi, 'Understanding the Technology Behind DeepFakes'(2018) Alan Zucconi Blog https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/ accessed 15 June 2020.
[3] See Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen "The State of Deepfakes: Landscape, Threats, and Impact" (2019) Deeptrace.
[4] Ibid.
[5] Will Knight, 'The US Military is Funding an Effort to Catch Deepfakes and Other AI Trickery'(2018) MIT TECH. REV.
[6] See Ian J GoodFellow and others, 'Semi-Supervised Learning with Generative Adversarial Networks' (2014) Cornell University https://arxiv.org/abs/1406.2661 accessed on 15 June 2020.

deepfake becomes extremely realistic and convincing.[7] Furthermore, in the near future it is plausible that GANs will be able to change a person's face, body and voice without the need for large amounts of data and images of the person portrayed. Thus, even a single photo might be enough to generate a highly realistic forgery.

Today, deepfakes creation software is easily accessible online and within everyone's reach.[8] This means that even the most technologically unskilled users are able to make a wide range of doctored videos. In fact, several tutorial videos on YouTube show how to use these technologies.[9] Any individual can therefore manipulate existing media and generate new fake content with relative ease, without necessarily understanding the complex mathematical and computational aspects behind deepfakes.

Moreover, it is important to mention that deepfakes emerged in late 2017 particularly in the field of pornography: faces of women celebrities were swapped on the bodies of pornographic actresses in hyper-realistic fake videos.[10] However, this non-consensual appropriation of individual identities has then expanded to include political actors whose image can be exploited by malevolent individuals to distort political ideas and influence public opinion.[11]

The BBC reported in November 2019 that research conducted by cyber-security company DeepTrace found that (at least) 14,698 deepfake videos are now online, compared with 7,964 in December 2018.[12] Therefore, the amount of these doctored videos circulating the internet has almost doubled in just twelve months. Combined with the phenomenal speed and broad reach of social media, deep fakes can spread extremely quickly and reach millions of people. Consequently, potential pitfalls on social structures may be highly disruptive. The global, immediate threat is already being perceives, considering the deceptive potential and the fact that anyone can access software to reshape a certain person in a digital video.

---

[7] John Donavan, 'Deepfake Videos Are Getting Scary Good' (2018) HowStuffWorks https://electronics.howstuffworks.com/future-tech/deepfake-videos-scary-good.htm accessed 18 June 2020.

[8] See Frederick Mostert, Henry Franks 'How to counter deepfakery in the eye of the digital deceiver' (2020) Financial Times https://www.ft.com/content/ea85476e-a665-11ea-92e2-cbd9b7e28ee6 accessed on 18 June 2020.

[9] See *ex multis:* Cinecom; "Deepfake Tutorial: A beginners Guide"(YouTube, 10 Dec 2019) https://www.youtube.com/results?search_query=how+to+create+a+deepfake+video accessed 15 June 2020.

[10] Douglas Harris, 'Deepfakes: False Pornography Is Here and the Law Cannot Protect You' (2018-2019) 17 Duke L & Tech Rev 99.

[11] Kietzmann, J., Lee, L. W., McCarthy, I. P.; Kietzmann, T. C. "Deepfakes: Trick or treat?"(2020) *Business Horizons* 63 (2) 135.

[12] *S*upra note 3.

2. The potential misuses of deepfakes

Deepfakes raise multiple legal concerns as they can be used to cause a wide range of serious harm which can be distinguished and classified according to whether they affect society in general, or individuals.[13] Before examining these potential damages, it is crucial to emphasize that deep fakes are a dangerous extension and progression of fake news. The US presidential elections in 2016 and the Brexit vote have already dramatically shown how campaigns of disinformation can manipulate public opinion, deceive the electorate and undermine confidence in institutions. At the time, deepfakes were not yet popular, but it is evident that "information diseases" were already widespread. Moreover, because of their deceptive nature, deepfakes are even more alarming and problematic than fake news. Therefore, if deepfakes are used with the deviant purpose of spreading false information, the damage to the democratic system might be disastrous. For instance, such AI-manipulated content could strongly interfere with the election campaigns. Thus a political candidate could be shown covering a misdeed or making racist comments right before an election or an important resolution.[14] The situation is even more aggravated, considering the ability of such digital content to undermine the natural human tendency to rely on visual and sound perception.[15] Until now, videos have been a relatively reliable source of information. But with the rise of deep fake, it is no longer enough to see to believe.[16] The impossibility of distinguishing authentic videos from fake ones may have the consequence of suggesting to suspect that all videos are considered manipulated and falsified.[17] This skepticism leads to a large-scale erosion of public faith in online information and content. On the other hand, ruthless political actors could use deepfakes to escape liability for certain speeches or actions by claiming that the authentic video is actually a deepfakes.[18]

In this way, trust in the political system becomes more difficult to establish and maintain.[19]

The dissemination of deliberately fabricated and misleading information also has dramatic implications in terms of public security. For example, experts have develop the hypothesis that ISIS or al-Qaeda terrorist groups might produce videos of American soldiers killing local civilians

[13] Ibid.

[14] Rebecca Green, 'Counterfeit Campaign Speech'(2019) Faculty Publications https://scholarship.law.wm.edu/facpubs/1923 accessed 18 June 2020.

[15] Elizabeth Caldera, 'Reject the Evidence of Your Eyes and Ears: Deepfakes and the Law of Virtual Replicants' (2019) 50 Seton Hall L Rev 177.

[16] Holly Kathleen Hall, 'Deepfake Videos: When Seeing Isn't Believing' (2018) 27 Cath U J L & Tech 51.

[17] This consequence is what B. Chensey and D. Citron have called 'Liar's Dividend' *see* Bobby Chesney and Danielle Citron 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 Calif L Rev 1785.

[18] Ibid.

[19] Ibid

on the battlefield in order to use them as means of propaganda.[20] Deepfakes would then be used to spark and justify waves of violence.

Moreover, such synthetic content could damage international relations and cause diplomatic incidents, as well as undermining specific military or intelligence operations. As this analysis suggests, the alarmingly realistic fake video raises serious concerns about how the whole democratic system works and what kind of information citizens will have access to. Such threats have a systemic dimension and contribute to the development of the so-called "perfect storm of disinformation".[21]

The creation of highly reliable false statements or facts also poses serious threats to individuals, facing new forms of exploitation, intimidation, personal sabotage, and enormous reputational damage. The first targets of the negative effects of deepfakes were the Hollywood actresses turned into the protagonists of fake tailor-made sex videos.[22] However, all individuals are potentially exposed to deepfake non-consensual pornography and the extensive damage caused by it, regardless of whether they are celebrities or not.[23] Indeed, if the victims of these events cannot prove that the video in question has been manipulated and is faked, they could suffer psychological damage and collateral consequences, for example in the workplace and in their private lives. Therefore, depending on the circumstances and the time of the spread of a deepfake, the effects for the targeted person could be dramatic. Furthermore, deepfakes could be used to profit from a person's image without his or her consent (i.e. false-endorsement cases). Another serious problem posed by image manipulation is identity theft. This danger is potentially enormous in terms of fraud, blackmail and cyber extortion. For instance, criminals could use this new technology to impersonate a chief executive officer in order to start a fraudulent transfer. This example clearly shows how also the financial and corporate world can be exposed to several cybersecurity attacks perpetrated through deepfakes (i.e. market and stock manipulation).[24]

As this overview shows, deepfakes are a global phenomenon with the potential of providing cybercriminals and other malicious online actors with tools to commit fraud, influence the general public, undermine national security and seriously harm individuals. Although the vast majority of

---

[20] Olivia Beavers, *'Washington fears new threat from 'deepfake' videos'*(2019) TheHill https://thehill.com/policy/national-security/426148-washington-fears-new-threat-from-deepfake-videos accessed 19 June 2020.

[21] Ibid.

[22] Caroline Drinnon, "*When Fame Takes Away the Right to Privacy in One's Body:* Revenge Porn and Tort Remedies for Public Figures"(2017) WM. & MARY J. WOMEN & L. 209, 211.

[23] Kristi Melville, 'Humiliated, Frightened and Paranoid: The Insidious Rise of Deep Fake Porn Videos' (2019) ABC News.

[24] Catherine Stupp, 'Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case' (2019) *Wall Street Journal* https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 accessed 19 June 2020.

these videos are created with pornographic content, technology is increasingly being used with a variety of sinister purposes to target populations.

3. The bright side of deepfake

Despite the multiple risks that could arise from the spread of a deepfake, this technology could have a silver lining.[25] In fact, deepfakes are not per se misleading and harmful but can be positively used in several sectors, such as entertainment and movie, educational and social media, and healthcare. Hence, synthetic media have proved to be highly innovative with some relevant benefits.[26]

Recently, the documentary filmmaker David France decided to expose the multiple civil rights violations in Chechnya and used the deepfakes to protect the identity of the protagonists.[27] Extensive post-production work has made it possible to change the faces of the 23 hunted individuals to avoid possible retaliation against them and at the same time maintain the high visual quality of the documentary.[28] Deepfakes therefore have a significant value in the creative multimedia context and could also transform the entertainment world: actors might lend their image rights for the creation of entirely synthetic content instead of physically acting in films. Another potential use in the film industry is the exploitation of deep fakes to produce movies with long dead actors.[29] There is also no need to mention the possible parody and satirical uses. Moreover, this technology could have positive results in the business market. Some companies have cut the cost of online training for their staff by producing synthetic videos, with computer-generated models that address users in their own language and call them by name.[30] Other possible beneficial applications of deepfakes concern the healthcare industry. For instance, researchers intend to use this technology to create highly realistic medical images to develop and train artificial intelligence to detect diseases.[31] The creation of synthetic images and data would also avoid some privacy issues and financial concerns that often occur in this field.[32] In such cases, the use of deep fakes serves legitimate purposes and must be preserved.

---

[25] See supra note 17 1753, 1769-1771.
[26] Baccarella, C. V., Wagner, T. F., Kietzmann, J. H., & McCarthy, 'Social media? It's serious! Understanding the dark side of social media'(2018) European Management Journal, 36(4) 431.
[27] Joshua Rothkopf 'Deepfake Technology Enters the Documentary World' (2020) New York Times.
[28] Ibid.
[29] Mika Westerlund 'The Emergence of Deepfake Technology: A Review' (2019) Tech. Inn. Management Review.
[30] See Jane Wakefield and Beth Timmins 'Could deepfakes be used to train office workers?' (2020) BBC News https://www.bbc.com/news/technology-51064933 accessed 20 June 2020.
[31] Jackie Snow, 'Deepfakes for good: Why researchers are using AI to fake health data', FAST COMPANY, https://www.fastcompany.com/90240746/deepfakes-for-good-why-researchers-are-using-ai-for- synthetic-health-data accessed 20 June 2020.
[32] Ibid.

CHAPTER II

How to regulate deepfakes without jeopardizing freedom of expression?

As has been discussed above, deepfake technology is not inherently malicious and its positive applications are undeniable. The use of this type of artificial intelligence is could then be framed as a particular form of constitutionally protected speech and expression. For this reason, it cannot be subject to an indiscriminate ban. In the United States, freedom of expression is highly protected by the First Amendment which prohibits Congress from making any law "abridging the freedom of speech", as public opinion is shaped through free and open discourse.[33] This fundamental right is enshrined and endowed with protections also under the European Union Charter of Human Rights[34] and the European Convention on Human Rights.[35]

Despite this strong legal protection, the freedom of expression could be restricted where the deepfakes contain intentionally false and deceptive images intended to create a "legally recognizable damage". Consequently, every legal remedy to combat the negative effects of deepfakes must go through a careful balancing test between freedom of speech and other fundamental rights.[36] In the next paragraphs, I will discuss how the courts and legal scholars in the United States and the European Union have understood and rationalised the protection of freedom of expression in terms of truth and falsehood. In particular, the following analysis shows that deep fakes could lead to a revision of traditional legal arguments on the regulation of information flows, particularly in the United States.

1. US Approach to Freedom of Expression

a. Marketplace of Ideas, the Counterspeech Doctrine and the Crucial Value of Free Speech

The Supreme Court of the United States has developed an extremely strict protection of freedom of expression since 1919, when Justice O.W. Holmes formulated the famous metaphor of the "marketplace of ideas": competition between different opinions is the best way to allow the truth

[33] U.S. Const Amend. I.
[34] Article 11 of EU Charter of Fundamental Rights.
[35] Article 10 of European Convention of Human Rights.
[36] Olivia Beavers "Washington fears new threat from 'deepfake' videos" (2019) The Hill https://thehill.com/policy/national-security/426148-washington-fears-new-threat-from-deepfake-videos accessed 1 July 2020.

to impose itself in a democratic society.[37] In his passionate dissenting opinion, Holmes stated: "the best test of truth is the power of the thought to get itself accepted in the competition of the market, and that truth is the only ground upon which their wishes safely can be carried out."[38] According to this theory, unpopular and even false ideas must be able to compete in the market place of ideas and, for this reason, public authorities must step aside by letting people discover and assess for themselves which ideas are worth following.

For this reason, as Supreme Court stated: "Under the First Amendment there is no such thing as a false idea. However pernicious an opinion may seem, we depend for its correction not on the conscience of judges and juries but on the competition of other ideas."[39]

The metaphor developed by Justice Holmes is inspired by the liberal economic notion of "marketplace" and has its roots in the Enlightenment philosophy, particularly, in John Milton's thought.[40] In the Areopagitica, Milton wrote: "Let [truth] and Falsehood grapple; who ever knew Truth put to the worse, in a free and open counter".[41] He criticizes press censorship on the grounds that it hampers the free flow of ideas that leads to the realization of an individual's intellectual faculties and the achievement of truth. Censorship could, in fact, influence the process of approaching the truth by preventing or curtailing the emergence of new information.[42]

Similarly, John Stuart Mill, also part of the same intellectual tradition, claimed that censorship "robb[ed] the human race."[43] Mill defends the protection of false speech and argues that ideological truth is not universally known, but must be discovered through discussion. Following Mill's reasoning, falsehood has a particular value precisely because it makes people investigate information further, and this leads to the discovery of the truth. If misleading ideas were censored, the truth would be inevitably compromised.[44]

Since the first appeal to the marketplace of ideas as a theory of free expression, it has been invoked countless times by the Supreme Court and federal judges to explain the scope of the First Amendment.[45] After Holmes introduced this legal theory, Justice Louis D. Brandeis implicitly recalled it and established the '*Counterspeech doctrine*' in the landmark case Whitney v.

---

[37] *Abrams v. United States*, 250 U.S. (1919) para. 616, 630.
[38] Ibid.
[39] *Gertz v. Robert Welch*, Inc. 418 U.S. (1974) para. 323, 339–40.
[40] John Milton, *Areopagitica* (1644; Jebb ed. Cambridge University Press, 1918) 58.
[41] Ibid.
[42] Ibid.
[43] John Stuart Mill, *On Liberty* (2d ed. 1863) 35-36.
[44] Ibid 41.
[45] See *ex multis: Bigelow v. Virginia* 421 U.S. (1975) 809; *Consolidated Edison Co. of New York v. Public Service Commission* 447 U.S. (1980) 530; *Board of Education v. Pico* 457 U.S. (1982) 853, 866-67; *Red Lion Broadcasting Co. v. FFC* 395 U.S. (1969) 367, 390; *Virginia v. Hicks* 521 U.S. (1997) 844, 855, *Reno v. American Civil Liberties Union* 521 U.S. (1997) 844; *McCreary County v. American Civil Liberties Union* 545 U.S. (2005) 844; *Randall v. Sorrell* 548 U.S. (2006) 230; *Walker v Sons of Confederate Veterans* 115 U.S. (2015) 2239.

California, reasoning that " if there be time to expose through discussion, the falsehoods and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence."[46] At the core of this other theory is the principle that "whenever 'more speech' could eliminate a feared injury, more speech is the constitutionally-mandated remedy."[47] Thus, the proper response to false and harmful speech is to add positive speech in the metaphorical marketplace of ideas.[48] Therefore, in Gerz v. Robert Welch, the Supreme Court, relying on counter-speech principle, stated that "the first remedy of any victim of defamation is self-help—using available opportunities to contradict the lie or correct the error and thereby to minimize its adverse impact on reputation".[49]

According to these two complementary theories, a free and provocative discussion allows the power of reason and truth to manifest itself. In this way, citizens can freely pronounce and exchange all kinds of artistic, scientific, commercial, even hateful and false expressions without fear of governmental coercion. It is worth mention that this approach is deeply rooted to its historical and social context.[50] The United States, in fact, has remained firmly attached to a liberal vision of fundamental rights which are perceived as a shield against abuses of State power. Accordingly, public authorities should refrain as far as possible from regulating the flow of information and setting limits to the free exchange of ideas.[51] In the light of this, the First Amendment does not allow the government, except in a few specific cases, to limit the categories of discourse since the effects of restrictive measures inhibit discussion on public issues.[52]

b. Potential Distortions and Cracks in the Marketplace of Ideas

The theories described so far have been defined as "the dominant rationale of freedom of expression" as they guide the Supreme Court's First Amendment jurisprudence.

However, despite the broad consensus reached, these principles are not sheer of critical aspects. In fact, some legal scholars have not shared the Supreme Court's full trust in the self-correcting

[46] *Whitney v. California* 274 U.S. 357 (1927) Brandeis J, concurring opinion.
[47] Laurence H. Tribe, American Constitutional Law 834 (2d ed. 1988).
[48] See David L. Hudson Jr. 'Counterspeech Doctrine'(2017)The First Amendment Encyclopedia https://www.mtsu.edu/first-amendment/article/940/counterspeech-doctrine accessed 1 July 2020.
[49] Supra note 39 323, 344.
[50] C. Edwin Baker, Human Liberty and Freedom of Speech (1989) 6-7.
[51] *New York Times v. Sullivan*, 314 U.S. para. 252, 270 (1964); *Linmark Associates Inc. v Townships of Willingboro* 431 U.S. 85 (1977); *Texas v. Johnson*, 491 U.S. para 397, 419-20 (1989); *Lorillard Tobacco Co. v. Reilly* 533 U.S. 525 (2001); *Virginia v. Black*, 538 U.S. 343, 358 (2003).
[52] *Police Department v. Mosley* 408 U.S. (1972) para. 95-96.

capacity of the market of ideas and, consequently, have highlighted some concerns about the process of transmission and reception of information in that metaphorical market.[53]

For instance, C. Edwin Baker argues that the "marketplace of ideas" theory fails every time it is applied because each individual perceives the world in his/her own way and, for this reason, processes information differently, depending on several factors, such as personal experience, race, socio-economic position, religion and other forms of socialization.[54] For this reason, the truth is never objective but depends instead on individuals' perception. Therefore, according to Baker, mere discussion is not an adequate tool to eliminate differences between people and it is often insufficient itself to determine the best choice between different ideas.[55]

Moreover, this metaphor ignores that people have imperfect abilities and often do not discern falsehood from truth, even if they believe they are able to do so. As a result, the validity of the decision-making process is compromised.[56] These problematic aspects are reflected and amplified throughout the Internet, which has become the new marketplace of ideas.[57] In fact, when online users are overwhelmed by information, they assume a particular attitude by looking for news and information that supports their previous ideas. Political scientists call this phenomenon the "filter bubble".[58] Thus, the pluralistic exchange of ideas is actually more limited, also considering the complex algorithms that dominate our online behavior and choices.[59] Indeed, the targeting and profiling techniques are used to create a specific universe of information for each individual user, isolating them from anything that is unrelated to their beliefs.[60] As a result, these "filter bubbles" affect the way in which individuals approach ideas and information, without guaranteeing a proper confrontation with other different views.[61] The concrete risk is that the market of ideas will turn out to be a mere illusion that does rarely lead to the emergence of truth.

It should be also noted that the validity of the "marketplace of ideas" paradigm depends on public access to the full range of information. Therefore, an essential assumption is the rational and

---

[53] See Frederick Schauer, "The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience" (2004) *Harvard Law Review* 117, no. 6 1765-809 doi:10.2307/4093304.

[54] Baker supra note 50, at 3-6.

[55] Ibid at 11-13.

[56] Paul H. Brietzke, 'How and Why the Marketplace of Ideas Fails' (1997) at 951, 962-63 Valparaiso University Law Review.

[57] *Reno v American Civil Liberties Union (ACLU)*, 521 U.S. 844 (1997) referring to the growth of the Internet as a "dramatic expansion of this new marketplace of ideas".

[58] The concept of filter bubbles was introduced by Eli Pariser in his revolutionary book entitled "The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think " (2012).

[59] Oreste Pollicino, 'Fake News, Internet and Metaphors (to Be Handled Carefully)' (2017) Italian Journal of Public Law 1.

[60] This dynamic emerged strongly during the US elections in 2016 and in the Brexit vote.

[61] Philip M. Napoli, 'What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble' (2018) 70 FED. COMM. L.J. 55, 57.

informed process for the selection of the truth.[62] While the digital world seems to make this process theoretically possible through the increased possibility of finding information and expressing thoughts, cultural prejudices and cognitive limitations, amplified by the above mentioned algorithms, hamper the efficient processing of knowledge.[63] Indeed, information sources frequently have an innate bias due to an intellectual, political, or market affiliation.[64]

Meanwhile, the Internet and its social media platforms play a key role in providing an efficient infrastructure for the rapid dissemination of ideological and politically false information to deceive individuals. The global nature of the "new" technologies and the fact that potentially every Internet user can spread and share false information on such platforms can lead to a careful review of the metaphor of the market for ideas, as these dynamics exponentially reinforce the urgent need to verify information sources in the digital age.

Moreover, some legal scholars argue that the principle of counter-speech is difficult to implement in a unfair market of ideas, where some people or groups in society have more power than others.[65] For instance, supporters of critical racial theory argue that "minorities are often denied access to the market for ideas to counter detrimental discourse".[66] Others argue that certain types of discourse are harmful to the point that counter-speech alone is not an adequate response, however persuasive or effective.[67] Nevertheless, First Amendment experts Robert Richards and Clay Calvert argue that "even if counter-speech is not always a perfect remedy, individuals and courts should seriously consider it as a solution. When used wisely, the counter-speech may prove to be a very effective solution for harmful or threatening expressions".[68]

c. The Supreme Court versus falsehood

Despite the weaknesses and critical aspects discussed above, the Supreme Court continues to use the metaphor of the marketplace of ideas and the counter-speech doctrine in First Amendment cases to justify the absence of a strong state interference in regulating online speech and even to protect the falsehood.[69] For instance, Justice Anthony Kennedy has cited Justice Brandeis' famous

---

[62] Claudio Lombardi, 'The Illusion of a 'Marketplace for Ideas' (2018) http://dx.doi.org/10.2139/ssrn.3104449 accessed 5 June 2020.
[63] Ibid.
[64] Ibid.
[65] Napoli supra note 61, 67-68.
[66] Mari J. Matsuda, 'Words that wound: critical race theory, assaultive speech, and the First Amendment' (1993) at 48.
[67] *Hustler Magazine, Inc. v. Falwell* 485 U.S. 46 (1988) at 52.
[68] Robert D. Richards & Clay Calvert, "Counterspeech 2000: A New Look at the Old Remedy for "Bad"Speech'(2000) B.Y.U. L. REv. 553, 585 (explaining that the best response to today's objectionable speech is counterspeech).
[69] J. Scott Harr, Kären M. Hess, Christine Hess Orthmann, Jonathon Kingsbury "Constitutional Law and the Criminal Justice System" (2015) Ch 5 p 146-147.

principle in his plurality opinion in United States v. Alvarez.[70] In that case, the Court struck down the constitutionality of the Stolen Valor Act, a law that broadly prohibited any false claim about military honors. Justice Kennedy would also seem to evoke Mill's theory when he writes "The remedy for speech that is false is speech that is true. This is the ordinary course in a free society. The response to the unreasoned is the rational; to the uninformed, the enlightened; to the straight-out lie, the simple truth."[71] This case has provided the Supreme Court with the opportunity to properly define the value of falsehood and the constitutionally admissible limits to freedom of expression.

First, the Judges underlined once again that the First Amendment prevents the government from restricting freedom of speech merely because of its content, ideas or message; consequently, any restriction based on the substance of the speech is presumed illegitimate unless the government proves its compliance with the Constitution. Justice Kennedy, in writing for the Court, characterized the Stolen Valor Act as a content-based restriction because it punished the perjury itself, without placing any emphasis on the purpose of the false statement and the possible pursuit of unfair advantage or profit. Kennedy explained: "The statute seeks to control and suppress all false statements on this one subject in almost limitless times and settings. And it does so entirely without regard to whether the lie was made for the purpose of material gain." For these reasons, the Stolen Valor Act did not fulfill the requirements to comply with the First Amendment. In fact, the possibility of criminalizing Alvarez's conduct - namely, falsely claiming that he had received a military honor - would have given the national authorities powers of censorship capable of jeopardizing the whole value of freedom of speech as a milestone of liberal democracy.

Secondly, the Supreme Court pointed out that the only restrictions allowed are limited to particular traditional categories of harmful expressions, including those involving incitement to unlawful conduct, defamation, obscenity, fraud. In these cases of 'low-value speech', there is a legally recognizable harm and government action is justified.[72] However, outside of these specific cases, no general exception exists to enable public authorities to exclude certain expressions from the First Amendment safeguards. Indeed, both Justice Kennedy's plurality opinion and Justice Breyer's concurrence argue that false statements - on "philosophy, religion, history, the social sciences, the arts and other matters" of public concern - represent a form of protected speech, necessary to safeguard democratic discourse and the marketplace of ideas.[73] Kennedy also explained " The

[70] Supra note 37.
[71] Ibid at 727 (plurality opinion).
[72] Alexander Tsesis, 'Categorizing Student Speech'(2018) 102 Minnesota Law Review at 1147, 1167 (discussing the Court's allowance of traditionally recognized forms of low- level speech).
[73] Supra note 37 at 724 (plurality opinion); at 730-731 (Breyer J. concurring).

Nation well knows that one of the costs of the First Amendment is that it protects the speech we detest as well as the speech we embrace. Though few might find respondent's statements anything but contemptible, his right to make those statements is protected by the Constitution's guarantee of freedom of speech and expression."[74]

Thus, the Supreme Court has deliberately equalized the protection of intentionally false statements to that traditionally granted to the most protected forms of speech. Despite the false content, in fact, these information, placed on the marketplace of ideas, still remain contributions to public discourse and cannot be censored.

By contrast, Justice Alito, in his dissenting opinion, argued that "this radical interpretation of the First Amendment is not supported by any precedent of this Court".[75] According to Justice Alito, false factual statements have no intrinsic First Amendment value, and should receive no protection "unless their prohibition would chill other expression that falls within the Amendment's scope".[76]

At this point, it is necessary to underline that the falsity-related cases prior to Alvarez outline a more complex picture than that presented so far. In New York Times v. Sullivan, for instance, the Court distinguished between intentional and unintentional lies.[77] In particular the justices suggested not to limit the accidental and involuntary falsehoods in the speech, reasoning that "erroneous statement is inevitable in free debate, and that it must be protected if the freedoms of expression are to have the 'breathing space' that they 'need (...) to survive."[78] On the other hand, the Court has indicated the standard of actual malice as an essential condition for the State to punish those who make false and defamatory statements against public figures. This provides a ground for compensation of the injured party. With regard to public figures, the same principle was confirmed in the criminal proceedings, in Garrison v. Louisiana.[79] Also in Hustler, the Court rationalized safeguarding unintentional errors in discourse arguing that "even though falsehoods have little value in and of themselves, they are 'nevertheless inevitable in free debate".[80]

However, in the same decision, the Court stated that false statements of fact could "interfere with the truth-seeking function of the marketplace of ideas."[81] The principle of the irrelevance of mere falsehood as such had already been stated in another leading case already mentioned, Gertz v.

---

[74] Ibid at 709, 729 (plurality opinion).
[75] Ibid at 746 (Alito J, dissenting).
[76] Ibid.
[77] *N.Y. Times Co. v. Sullivan*, 376 U.S. (1964). at 270- 271.
[78] Ibid.
[79] *Garrison v. Louisiana*, 379 U.S. (1964) 64, 75 (acknowledging that in defamation cases, "the knowingly false statement and the false statement made with reckless disregard of the truth, do not enjoy constitutional protection").
[80] *Hustler Magazine and Larry C. Flynt, Petitioners v. Jerry Falwell*, 485 U.S. (1988) 46.
[81] Ibid at 46, 52.

Robert Welch.[82] On that occasion, the Supreme Court ruled that each State is free to provide for a specific standard of liability for defamation, as long as it satisfies the minimum requirement of guilt. Thus, according to the Supreme Court, applying a strict liability for the defamation offence would be contrary to the First Amendment. In this way, as pointed out by Justice Breyer's concurring opinion in Alvarez, the concern that an individual in good faith might be accidentally liable is minor. In addition, in Gertz, justices made a distinction between false ideas and false statements of fact, ruling that "there is no constitutional value in false statements of fact. Neither the intentional lie nor the careless error materially advances society's interest in 'uninhibited, robust, and wide-open' debate".[83] In other First Amendment cases, however, justices pointed out that "untruthful speech, commercial or otherwise, has never been protected for its own sake."[84]

The Supreme Court's judgments are therefore not always consistent and depend on the interpretation given to the marketplace of ideas. Decisions in which the Court does not protect intentionally false and misleading information reveal concerns about possible distortions in the marketplace for ideas. In Alvarez, on the other hand, Justice Kennedy refers to the notion of marketplace precisely to protect the falsehood and strongly reaffirms the theory introduced by Holmes: discussion over ideas, as well as competition between different products, will lead to the best result in terms of truth. Kennedy insists that the limitation of freedom of speech by the government, through the introduction of incriminating speech, does not help to reveal and isolate the falsehood, but on the contrary only makes it more difficult. Kennedy stresses, then, that only a weak society needs protection or intervention by the public authorities to protect or preserve the truth.

d. Deepfakes rocking the marketplace of ideas at its foundations

The crucial framework just described constitutes the most relevant precedent regarding the falsity which could broadly support protections for deepfakes.

In Alvarez, the Supreme Court has in fact extended the protection of the First Amendment to intentionally false information such as those that could potentially arise from such AI-generated content. In case of government intervention against deep fakes, any restrictive measure would be exclusively based on the misleading video content and, following the reasoning of the Court in Alvarez, would not pass the constitutional scrutiny.[85] However, Justice Kennedy has pointed out

[82] note supra 39.
[83] Ibid 323, 339.
[84] *Virginia State Pharmacy Board v. Virginia Citizens Consumer Council* 425 U.S. (1976) 748, 770.
[85] Supra note 37 at 709, 723.

that "[w]here false claims are made to effect a fraud or secure moneys or other valuable considerations (…) it is well established that the Government may restrict speech without affronting the First Amendment."[86] In some cases, the use of techniques that allow image manipulation and the creation of forgery content could fall into the category of unprotected speeches, considering the risks and potential threats both for society and for the individual.[87]

Consequently, in relation to deepfakes, scholars and jurists should revisit very carefully both the theory of the marketplace of ideas and counter-speech doctrine. In fact, the fundamental assumption underlying these theories is inevitably undermined, since one of the most disruptive effects of deep fakes is to blur the line between truth and falsity and to compromise the ability of individuals to critically evaluate information. Consequently, it can be provocatively argued that the current information ecosystem weakens the validity of both the marketplace for ideas and the counter-speech doctrine because it gives too much power to those who spread disinformation, fake news and deep fakes.

The impact on democracy is definitely negative. Accepting the circulation of false and highly deceptive content could erode and affect the free and open debate in the marketplace of ideas that the First Amendment aims to preserve. As Philip Napoli states, the "algorithmic market of ideas" is a realm where "reliance on the counter-speech is increasingly ineffectual and potentially damaging to democracy."[88] Some other experts have in fact underlined America's unconvincing response to combat harmful propaganda in 2016, recognizing that "it is not enough to try to counter a firehose of falsehood with a squirt gun of truth".[89]

The need for action is clear and the US seems well aware of this, since a number of laws dealing specifically with deep fakes have been introduced at both federal and state level. These new legal solutions that will be discussed below, are important to understand the extent to which US lawmakers can limit the use and spread of harmful deepfakes without compromising freedom of expression (for example, by providing exceptions for legitimate uses of deep fakes such as satire).

---

[86] Ibid.

[87] See Danielle K. Citron 'Hate Crimes in Cyberspace' Harvard University Press (2014) at 199-218.

[88] Napoli supra note 61 at 97.

[89] William Courtney & Christopher Paul, "Firehose of Falsehoods:Russian propaganda is pervasive, and America is behind the power curve in countering It" U.S. NEWS & WORLD REP (September, 2016) https://www.usnews.com/opinion/articles/2016-09-09/putins-propaganda-network-is-vast-and-us-needs-new-tools-to-counter-it accessed 5 June 2020.

2. European Union Approach to Freedom of Expression

a. An overview of EU freedom of expression and the main differences with US legal system

Despite the common liberal background, the American approach and the European one recognize a different level of protection for freedom of expression.[90] In the United States, as outlined above, freedom of expression has been broadly rationalised, relying on marketplace of ideas theory and counter-speech doctrine. On the other hand, European jurists have theorized freedom of expression differently, giving it a more limited scope and tolerating more the possibility of state regulatory intervention in information flows.

In fact, while the US Supreme Court insists on the non-interference of public authorities with this fundamental right and assumes that content-regulation does not bring it any benefit, European courts deem it necessary to carefully balance freedom of expression with other rights, such as the right to privacy and citizens' reputations.

For this reason, they justify the lawfulness of remedies that may restrict in some way the exercise of free speech. This rationale is based on paragraph 2 of Article 10 of the European Convention on Human Rights which, referring to "duties and responsibilities", opens the possibility for public authorities to interfere with this freedom by way of “formalities, conditions, restrictions and even penalties”. After the solemn affirmation of freedom of speech in paragraph 1, in fact, the rule provides for three requirements that any limitation of the freedom in question must comply with (so-called 'triple test').[91] In particular, the limitations must be prescribed by law, must be proportionate ("necessary in a democratic society") and must be aimed at the achievement of one of the objectives explicitly provided for in the same second paragraph of article 10 including national security, public safety, and the "protection of health or morals”.

The difference with the First Amendment is clear. This milestone of American democracy is set in absolute and solemn terms, simply by establishing a ban on public authorities.

The potential restrictions on freedom of expression, as discussed above, are therefore not directly laid down in the words of the First Amendment but have been developed by case law, according to restrictive criteria. The absence of such an explicit list of exceptions in the Constitutional document can help explain the divergence of the rationales of these two legal systems.

---

[90] See S. Choudry, ‘The Migration of Constitutional Ideas’ Cambridge, 2007 Ch 6.

[91] Dirk Voorhoof, “The Right to Freedom of Expression and Information under the European Human Rights System: Towards a more Transparent Democratic Society” EUI Working Paper RSCAS 2014/12.

Moreover, it is necessary to underline how, in Europe, doctrine and jurisprudence have outlined a complex and articulated content for the freedom of speech, which includes a multitude of declinations.[92]Article 10 in fact does not only protect the 'speaker' but also the 'listener', who has the right to receive information.

Thus, the European Convention on Human Rights, together with the Constitutions of the EU Member States, recognizes the right to information as a qualified part of freedom of speech, and outlines this freedom into two different aspects: the active one (right of access to public information)[93] and the passive one (right to receive information).[94] This elaboration has no equivalent within the US model and it is quite clear that the existence of an explicit or, in any case, consolidated constitutional guarantee of freedom of information supports mechanisms that can "purify" the web from contents that are incapable of expressing a valid information contribution as completely false or lack of qualification or verification.

Then, if the right to receive information is protected, inevitably the need for uncontaminated information, aimed at a virtuous formation of public opinion, will have to be safeguarded.[95] In many instances, in fact, the courts have communicated that information is a public good and as such should be protected.[96] This argument potentially supports a view that identifies a limit to the disruptive capacity of the deepfake problem, which exponentially increases the levels of disinformation. If, indeed, freedom of expression is functional to the satisfaction of the information needs of the citizens, and therefore, indirectly, to the formation of public opinion and the correct functioning of democracy, the recognition of the right to be informed can be translated into the provision of a "right to correct information". Such recognition, both explicit and implicit, provides the grounds for believing that there is a strong basis in European constitutionalism to counter the spread of disinformation, false news and even falsifications.

---

[92] Ibid.

[93] *Timpul Info-Magazine and Anghel v Moldova* App. no. 42864/05 ECtHR (27 January 2007) § 31.

[94] *Társaság a Szabadságjogokért v Hungary*, App no 37374/05 ECtHR (2009), § 26. In this case, the Court broadly interpreted Article 10 and implicitly recognized the right of access to official documents.

[95] *Steel and Morris v. UK* App No 68416/01 ECtHR (2005) §79; Hertel v Switzerland No 25181/94 ECtHR (1998) § 46.

[96] *Aquilina and Other v. Malta* App. No 28040/08, ECtHR (2011)*; Kobenter and Standard Verlagz GMBH v. Austria* App. no. 60899/00 ECtHR (2007).

b. Balancing freedom of expression and other fundamental rights: the core of European approach

If the European Union cares about the safeguarding of the proper information available to individuals,[97] public authorities and lawmakers should not be deceived and believe that the mere excess of information sources is sufficient to meet that objective and thus ensure pluralism.

It has already been pointed out that the digital information environment does not guarantee a correct flow of information, given the various pathologies that afflict it (filter bubbles, information cascades, cognitive bias).[98] Moreover, the speed of dissemination of online content must also be taken into account. In Delphi vs Estonia, the European Court of Human Rights explained how "defamatory and other types of clearly unlawful speech, including hate speech and speech inciting violence, can be disseminated like never before, worldwide, in a matter of seconds, and sometimes remain persistently available online".[99] This reasoning highlights the unique ability of online communication to damage reputation and invade individuals' privacy. This problematic situation could be amplified exponentially with deepfakes. In Delphi, the Grand Chamber has confirmed the trend described so far, inclined to more favorably validate possible limitations of online freedom of expression. Thus, the Court held that imposing a liability on forum provider (Delphi) for defamatory third-party comments did not constitute a disproportionate restriction on the applicant company's right to freedom of expression.[100] The Court's concrete assessment took into account both the inadequacy of the measures taken by Delphi company under the notice-and-takedown system it created[101] and the moderate sanction imposed on it. This decision can be justified by the need to balance freedom of expression on the Internet with the rights relating to the honour and reputation of the libeled person. However, it is necessary to specify that personality rights are not the only rights that European lawmakers and courts struggle to protect in this difficult balance with freedom of expression. For instance, in Mouvement Raëlien Suisse v. Switzerland, the ECtHR judges decided that interference by the Swiss government with freedom of expression was justified by the legitimate concern "to prevent crime, to protect health or morals and to protect the rights of others".[102] In the court's view, measures taken by the Swiss authorities to proscribe the display of posters (contrary to public order and immoral) were not in violation of Article 10 as justified by sufficient public-interest grounds. In

[97] This is reflected in the various policies to combat disinformation in the digital single market https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation accessed 6 June 2020.
[98] supra note 61; consider also the discussion at paragraph 1(a) in this chapter.
[99] *Delfi AS V. Estonia* App. No 64569/09 ECtHR (2015) at § 110.
[100] Ibid § 162.
[101] The defamatory comments remained accessible for about six weeks on the forum provider.
[102] *Mouvement Raëlien Suisse v. Switzerland App*. No. 16354/06 Eur. Ct. H.R. (2012) § 54-55.

that case, the Court drew attention to paragraph 2 of Article 10 and thus to the margin of appreciation available to the Contracting States in their assessment of need for and extent of an interference in the freedom of expression.[103] In particular, "the breadth of such a margin of appreciation varies depending on a number of factors, among which the type of speech at issue is of particular importance."[104] In addition to this, there are other strong reasons in the European legal system for introducing filtering mechanisms for false, synthetic and manipulated content. For example, the fact that the EU public authorities have a gatekeeping role would encourage greater control.[105]

In Europe, therefore, freedom of expression is seen as a right subject to balancing logic and therefore, at times, yielding to the emergence of overriding protection of different rights. Despite this, the judges of the ECHR emphasized that freedom of expression is a fundamental pillar of our democracy. In the Handyside case, the Court has described freedom of expression as "one of the essential foundations of a democratic society, one of the basic conditions for its progress and for the development of every man".[106] In that judgment, the Court found that a ban imposed by the British government on a book called Little Red School Book under the Obscene Publications Act was in accordance with the exception laid down in Article 10, §2 regarding protection of morals. Despite this decision, the Court also specified that the protection contained in Article 10 is "applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population".[107] In a subsequent passage, the ECHR underlined how the breadth of this area of protection meets the "requirements of pluralism, tolerance and broadmindedness without with there is no democratic society".[108] Also in Steel and Morris v. U.K., the court noted that there is " a strong public interest in enabling such groups and individuals (...) to contribute to the public debate by disseminating information and ideas on matters of general public interest".[109] Similarly, in Hertel v. Switzerland, the Court pointed out that "freedom of expression constitutes one of the essential foundations of society and one of the basic conditions for its progress".[110] It has also reiterated that free expression includes unpopular information, as well as content that can "offend, shock or disturb." The rulings discussed so far significantly expand the boundaries of freedom of expression, stating that it is not the level of social acceptance of a content that justifies its access to the guarantees of Article 10. This case law of the European Court of Human Rights bears similarities with U.S. legal system and the marketplace of

[103] Ibid § 59 citing *Tammer v. Estonia*, no. 41205/98, ECHR (2001) § 60.
[104] Ibid § 61.
[105] This trend emerges above all in the European regulations on privacy and copyright.
[106] *Handyside v. the United Kingdom*, App. No. (1976) § 49.
[107] Ibid.
[108] Ibid.
[109] *Steel and Morris v. UK* supra note 95.
[110] Ibid § 46.

ideas theory. Furthermore, it would seem to support the assumption that the broader and richer the landscape of ideas and opinions, the more the democratic character of a society will benefit.[111] However, it is necessary to gauge these decisions with the variety of the features of the malicious deepfakes, which lack a specific informational value and have effects contrary to the aims of freedom of expression, by spreading unrealistic harmful content to the community. The European courts must therefore ask themselves whether a certain online content, completely devoid of any social utility because it is false and accurately counterfeited, still deserves to be circulated or can be reported as false and possibly removed. In the light of the development of a right to information, filtering systems could be considered legitimate insofar as they respect a balance between the fundamental rights at stake, in accordance with an approach in line with the principle of proportionality. European public authorities should therefore intervene to regulate this phenomenon, with the aim of eliminating only those harmful deepfakes that undermine freedom of speech itself and also other rights.

The discussion so far provides crucial insights into how the European courts could deal with deep fakes in future cases. It is evident that the ECtHR is clearly concerned about the risks and damage that internet content and communications could cause to the exercise of human rights and freedoms. Moreover, given the lack of a strong duty of refraining on the part of the public authorities, the European legal ground is conceptually more inclined to support and implement forms of deepfakes prevention than the United States. Despite this, the EU has not yet adopted specific restrictive measures to tackle deep fakes and this could increase the risks caused by their misuse.

[111] see supra note 91.

CHAPTER III

THE CURRENT TOOLKIT: EXISTING LEGAL REMEDIES IN US AND EU SYSTEMS TO TACKLE MALICIOUS DEEPFAKES

The legal framework must necessarily evolve in order to successfully react to threats and damage arising from the malicious use of artificial intelligence. While the United States has recently started to regulate deepfakes in the context of both elections and pornography, the EU's actions in this area are almost entirely absent. Consequently, I argue for stronger law enforcement against deepfake technology to discourage the creation and dissemination of this harmful content.

1. United States Legal Framework

In the United States, growing social alarm about the harmful implications of AI-manipulated digital content is leading to increasingly aggressive anti-deepfake legislative action to address the malicious uses of them, particularly in the run-up to the 2020 presidential elections. As discussed in the second chapter, any state- or federal-level laws adopted to regulate deepfakes could interfere with freedom of expression and could therefore be subject to the First Amendment scrutiny. Accordingly, the new laws must be drafted carefully and narrowly to avoid infringing the right to freedom of expression. In the following paragraphs, I will give an overview of the specific US anti-deepfake laws and other existing legal remedies that could potentially be used, underlining the most problematic aspects of the latter.

a. Specific Anti-Deepfake Legislation

1. State law

In September 2019, Texas was the first state to specifically criminalize the use of deepfakes in the context of political elections. Texas Senate Bill 751 (SB751) amended the state's Election Code to ban deepfake videos created "with intent to injure a candidate or influence the result of an election" and which are "published and distributed within 30 days of an election".[112] This law does not provide for disclosure exceptions, and holds deepfake publishers liable for violations.[113] This could potentially allow social media platforms like Facebook, Twitter or YouTube to be held liable for

[112] Texas Senate Bill 751; 'Relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election.' (Texas) https://legiscan.com/TX/text/SB751/id/1902830 accessed 20 July 2020.
[113] Carolyn Toto 'Protecting Elections: Regulating Deepfakes in Politics' (2020) Pillsbury - Internet & Social Media Law Blog https://www.jdsupra.com/legalnews/protecting-elections-regulating-39567/ accessed 8 August 2020.

deepfakes posted on their sites, although such liability would seem to be preempted by federal law.[114] Such a strict regime may not survive the constitutional challenge. Therefore, the most likely consequence is that the Supreme Court will strike down this law both because of its possible conflict with Section 230 of the Communication Decency Act and because of the potential violation of the First Amendment given the lack of exceptions. This new Texas law closely follows in the footsteps of the State of Virginia, which in July 2019 became the first jurisdiction to legislate against pornographic deepfakes. Virginia has in fact expanded laws criminalizing the unauthorized distribution of sexually explicit material with malicious intent, to include the illegal distribution of non-consensual "falsely created" pornographic images and videos.[115] The consequences of such criminal offense are severe: imprisonment for up to 12 months and a fine of up to $2,500.[116] In contrast to the Texas law examined above, this law specifically exempts from liability Internet service providers who provide access to computers to users who commit such criminal acts.

In the wake of these new legal innovations, California has also passed two legislative acts to address the deepfakes issue.[117] Firstly, Assembly Bill No. 602 seeks to regulate the use of pornographic deepfakes by prohibiting the creation and dissemination of any “sexually explicit material” not authorized by the person concerned: this very far-reaching provision explicitly states that such conduct constitutes a crime regardless of its harmful intent, since it is intrinsically contrary to the fundamental rights of citizens.[118] Secondly, the Assembly Bill No. 730 does not directly address the deep fake per se, but prohibit the distribution of “materially deceptive audio or visual media” of a political candidate “with the intent to injure the candidate’s reputation or to deceive a voter”.[119] Unlike Texas and Virginia, California has not criminalized the creation or sharing of forgeries about election candidates. This law merely allows the election candidate portrayed in a deceptive photo or video to bring a civil action against the person or entity that distributed such content in the 60 days prior to the election. This piece of legislation is effective until 2023 and defines as “materially deceptive audio or visual media” any altered and manipulated content that is perceived as authentic by a reasonable person.[120] Essentially, the reasonable person must have a "fundamentally different

---

[114] 47 U.S. Code § 230 - Protection for private blocking and screening of offensive material.

[115] House Bill No. 2678 ‘Unlawful dissemination or sale of images of another; penalty’ https://lis.virginia.gov/cgi-bin/legp604.exe?191+ful+HB2678S1&191+ful+HB2678S1 accessed 20 July 2020.

[116] Code of Virginia § 18.2-11. ‘Punishment for conviction of misdemeanor’.

[117] Will Fisher ‘California's governor signed new deepfake laws for politics and porn, but experts say they threaten free speech’ (2019) Business Insider https://www.businessinsider.com/california-deepfake-laws-politics-porn-free-speech-privacy-experts-2019-10?r=US&IR=T accessed 20 July 2020.

[118] Assembly Bill - 602 Depiction of individual using digital or electronic technology: sexually explicit material: cause of action (California), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201920200AB602 accessed 20 July 2020.

[119] Assembly Bill -730 Elections: deceptive audio or visual media (California) https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 accessed 20 July 2020.

[120] Ibid.

understanding or impression" of the content in question than that person would have if he or she heard or saw the original, unaltered version of that content. In addition, this Californian law contains some significant exceptions. For example, radio and television stations paid to broadcast materially misleading media are exempted from the prohibition, regardless of whether the broadcasting station makes a statement regarding the lack of authenticity of the content. Conversely, images or videos published by the media as part of bona fide news, websites and regularly published periodicals can only benefit from the exception if the distribution of that content is clearly accompanied by an acknowledgement that the image or video is inaccurate or that there are doubts about its authenticity. Furthermore, the prohibition does not apply to those media that constitutes satire or parody.[121]

This Californian law seems to be a more balanced and narrowly tailored measure than those adopted in Texas and Virginia. Nevertheless, authoritative experts are concerned about the impact that this law could have on parody, political commentary and more broadly on freedom of expression. [122] On the other hand, some argue that Assembly Bill No 703 is too feeble, claiming that the exceptions laid down are too ambiguously worded and this provides a loophole for malicious actors. In addition, it will not be easy to prove actual malice in deepfakes, as it is often difficult to find clear and convincing evidence in this context. Given the high burden of proof, it is likely that a lengthy judicial review process will lead to a persistent proliferation of deepfakes.[123]

Other states have also proposed, but not yet approved, laws that seeks to prevent the harmful uses of deepfakes in certain contexts. For example, the legislation introduced in Massachusetts criminalises the use of deepfakes for conduct that is already "criminal or tortious", effectively making the use of a deepfake illegal in conjunction with the commission of other crimes.

While, in New York City, lawmakers have not yet considered a particular deepfake legal remedy, they are currently considering a law that would update the the right of publicity to protect an individual's digital image for 40 years after his or her death.[124] The bill would also create a register for surviving family members to record control of the digital image of the deceased individual. Furthermore, this piece of legislation contains strict prohibitions against the distribution of digitally created, sexually explicit works without clear and written consent from the person depicted.

---

[121] Ibid.

[122] For instance, David Greene, Electronic Frontier Foundation's Civil Liberties Director, argued that new legislation is unnecessary to counter deepfakes as a number of existing laws are sufficient. Contrary to this, the chapter will highlight how existing legal instruments can be helpful only to a certain extent. https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them accessed 23 July 2020.

[123] Brandie M. Nonnecke "Opinion: California's Anti-Deepfake Law Is Far Too Feeble" (2019) Wired https://www.wired.com/story/opinion-californias-anti-deepfake-law-is-far-too-feeble/ accessed 23 July 2020.

[124] Assembly Bill A5605C Establishes the right of publicity and provides for a private right of action for unlawful dissemination or publication of a sexually explicit depiction of an individual (New York) https://www.nysenate.gov/legislation/bills/2019/A5605 accessed 23 July 2020.

The aforementioned laws directly address concerns about both political disinformation and non-consensual pornography by establishing civil and criminal remedies. This legislative trend is likely to flourish in other states as well, taking into account the federal initiatives in this area that will be outlined below.

2. Federal Law

Congress is currently considering legislation to facilitate research and analysis of deepfakes media and the technologies used to generate them.[125] For example, the National Defense Authorization Act for Fiscal Year 2020 (NDAA) is a new political law signed by Trump that explicitly addresses the risks of deepfakes by imposing reporting requirements.[126] In particular, Section 5709 requires the Director of National Intelligence to submit to the Congressional Intelligence Committees an annual report describing the potential national security threats of deepfakes and their actual or potential use by foreign governments "to spread disinformation or engage in other malign activities". The same Section requires the Director of National Intelligence to notify Congress whenever there are substantial grounds for believing that a foreign entity has or is deploying deepfakes "aimed at the election or domestic political process of the United States." In addition, the NDAA establishes the "Deepfakes Prize" competition to develop and commercialize new technologies that allow the automatic detection of deepfakes, with awards of up to $5 million. The adoption of this law demonstrates that federal legislators recognize not only the serious consequences of deepfakes in the political context, but also the inadequacy of existing technological instruments to address them. Nevertheless, this piece of legislation does not provide for civil or criminal sanctions for the creation of such manipulated videos.[127]

Another law with the same research objectives as the NDAA is The Identifying Outputs of Generative Adversarial Networks ("IOGAN") Act which is now under Senate revision.[128] This bill would require the Director of the National Science Foundation (NSF) to support research on generative adversarial networks (GANs) and technical tools for verifying the authenticity of information and identifying manipulated media. In addition, public understanding of deepfakes would also be the subject of

---

[125] Matthew F. Ferraro, Jason C. Chipman, and Stephen W. Preston, 'The Federal "Deepfakes" Law' (2020) Journal of Robotics, Artificial Intelligence & Law Volume 3, No. 4.

[126] S.1790 116th Congress (2019-2020) National Defense Authorization Act for Fiscal Year 2020 https://www.congress.gov/bill/116th-congress/senate-bill/1790/text?q=%7B%22search%22%3A%5B%22deepfakes%22%5D%7D&r=14&s=1 accessed 23 July 2020.

[127] Ibid.

[128] H.R. 4355, 116th Cong.- Identifying Outputs of Generative Adversarial Networks Act (2019), *available at* https://www.congress.gov/bill/116th-congress/house-bill/4355/actions?KWICView=false accessed 23 July 2020.

research, as well as the identification of best practices for public education.[129] If this law passes, the directors of the NSF are required to submit a report to Congress on the results obtained with eventual policy recommendations that could ease and strengthen communication and coordination between the private sector, the NSF and the relevant federal agencies through the implementation of new strategies to identify and combat deepfakes.[130]

It is also worth mentioning the Deepfake Report Act of 2019, which is currently still pending in the House of Representatives.[131] According to this bill, the Department of Homeland Security should issue a report on potential malicious uses of deepfake technology (defined as a "digital content forgery technology"), describing the possible implications in terms of fraud and violation of federal civil rights. In addition, methods to detect and combat such forgeries will need to be assessed. The report should be submitted within one year of the enactment of the law and then every year for five years thereafter.

The legislation discussed so far is aimed primarily at improving research and understanding of deepfakes, without providing any significant protection against the damage caused by some of them. However, there are some bills currently being considered by Congress in order to tackle the abuses committed through deepfakes. In June of 2019, Representative Yvette D. Clarke proposed a controversial act that would require a creator of a deep fake to label the media with an “irremovable digital watermark”, disclosing that media has been manipulated.[132] Any failure to make the required disclosures, or any removal of the disclosures, would result in a civil penalty of up to $150,000. In addition, legislation would impose a criminal penalty of up to five years' imprisonment on any person who knowingly omits to label the deep fake or knowingly removes the watermark, if the intent is to humiliate, incite violence, interfere with an official proceeding, including elections, or otherwise engage in fraud-related criminal conduct.[133]

Notably, the act entitles the person or entity whose likeness is used in a harmful deepfake to bring a civil action, as long as the deepfake does not contain an appropriate disclosure label.

Moreover, this law establishes a task force within the Department of Homeland Security in order to strengthen the U.S. government's efforts to combat the national security implications of deep fakes, in addition to the development of technologies for the detection of deep fakes. These technologies should be made available to private sector Internet platforms, including social networks.

---

[129] Ibid.
[130] Ibid.
[131] S. 2065, 116th Cong. Deepfake Report Act of 2019 (2019), https://www.congress.gov/bill/116th-congress/senate-bill/2065/actions?KWICView=false accessed 25 July 2020.
[132] H.R. 3230 - Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019 https://www.congress.gov/bill/116th-congress/house-bill/3230/text accessed 25 July 2020.
[133] Ibid.

Overall, this law would seem to be the right answer to this emerging technology as it would significantly limit the diffusion of fake content online through restrictions on harmful deepfakes video. Despite this, there have been several criticisms. Some argue that the act is formulated in such broad terms that it could probably affect all fake videos and images that recreate a person, even those that constitute satire and are therefore protected by the First Amendment.[134] Some others have highlighted the "attribution problem": those who disseminate malicious or deceptive content online are likely to use anonymizing technologies. Consequently, the creator of the incriminated deepfake circumvents detection and therefore avoids liability under the discussed law.[135]

Beyond these problematic aspects, it is remarkable that the DEEPFAKE Accountability Act makes it clear which uses of deepfakes are against the law and, through this, provides a legal basis for both victims and law enforcement authorities to fight against malicious actors.

Finally, the House of Rappresentative have recently introduced a bill prohibiting individuals, political committees, and other entities "from distributing with actual malice any materially deceptive audio or visual media of a candidate within 60 days of a federal election with the intent to (1) injure the candidate's reputation, or (2) deceive a voter into voting for or against the candidate."[136] Thus, the bill establishes a new criminal offense related to the distribution of materially deceptive audio or visual media prior to a federal election. A violator is subject to a fine, up to 5 years in prison, or both. Although this act is still at an early legislative stage, it shows the US Government's concerns about deepfakes' political threats and its willingness to preserve future presidential elections..

b. Shifting the viewpoint: other potential legal remedies

Some experts have argued that the malicious use of deep fakes can be effectively prevented through a combination of intellectual property and state tort law.[137] In some cases, victims may bring a civil action claiming copyright infringement, defamation, and false advertising. Moreover, the right of publicity may also have been used to deal with the deviant uses of deep fakes. However, some of these remedies fail to provide real legal protection.[138]

For example, the victim could sue the deepfake creator for copyright infringement, claiming both monetary compensation for the exploitation of copyrighted content and the removal of the harmful

[134] Zachary Schapiro,'DEEP FAKES Accountability Act: Overbroad and Ineffective (April, 2020) Intell. Prop. & Tech F
[135] See Devin Coldewey, "DEEPFAKES Accountability Act would impose unenforceable rules-but it's a start" https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/ accessed 25 July 2020.
[136] H.R.6088 - Deepfakes in Federal Elections Prohibition Act.
[137] see supra note 122.
[138] For a comprensive review of the issue see supra note 10 (Harris Douglas).

deepfake through the notice and takedown mechanism. However, this action primarily assumes that the victim is the actual copyright owner. Moreover, the chances of success in these copyright infringement cases are considerably limited as they also depend on the court's application of the fair use doctrine.[139] Indeed, although the victim enjoys certain exclusive rights to the images or videos used in a deepfake, the court could protect the deepfake as a fair use by allowing the unlicensed use of copyrighted works for educational, critical, artistic and other expressive purposes.[140] In determining whether use made with a deepfake is a fair use, the court will have to consider certain factors such as "the purpose and character of the use, the nature of the copyrighted work, the amount and substantiality of the portion used in relation to the copyrighted work as a whole and the effect of the use upon the potential market for or value of the copyrighted work".[141]

Under these factors, the Supreme Court has brought the concept of transformative use.[142] When the new work is "transformative" or "adds something new, with an ulterior purpose or a different character", the doctrine of fair use can be extended to protect the work.[143] Thus, a deep fake creator could rely on this defense, since deep fakes are transformative works par excellence.

This first analysis suggests that copyright law is not an effective safeguard for those subject victimized by non-consensual deepfakes, especially when courts extend the doctrine of fair use.

A further form of limited protection is the right to publicity in civil matters. This right allows compensation for non-consensual use of the victim's likeness for commercial purposes.[144] Thus, the right of publicity strongly protects the economic value of a person's image or likeness and therefore turns out to be a feeble remedy for non-famous plantiffs.[145] Moreover, the financial gain element is rarely the ultimate goal of the deepfake creator. So, only in some cases (e.g. false endorsement), such an approach could prove useful to prevent abuse.

Other privacy-related torts may seem relevant at first glance, but they also remain of limited use. For example, "public disclosure of private facts" protects individuals from disclosures of intimate details of their lives that are not generally known and which would offend the average person.[146] However, the person portrayed in the manipulated content did not actually perform those acts and therefore the creation of the deepfake does not constitute a ‘revelation of a true intimate detail of that person's

---

[139] Tom Kulik, ‘Faking It: Why Deepfakes Pose Specific Challenges Under Copyright & Privacy Laws’ (2019) Above the Law https://abovethelaw.com/2019/07/faking-it-why-deepfakes-pose-specific-challenges-under-copyright-privacy-laws/?rf=1 accessed 25 July 2020.

[140] The Digital Millennium Copyright Act, 17 USC §107 (1998).

[141] Ibid.

[142] *Campbell v. Acuff-Rose Music Inc.*, 510 U.S. (1994) 569, 579.

[143] Ibid.

[144] J. Thomas McCarthy, Roger E. Schechter, The Rights of Publicity and Privacy, § 1:2, Clark Boardman Callaghan (2d Ed., 2020).

[145] Ibid.

[146] Restatement (Second) of Torts § 652B (AM. LAW INST. 1977).

life'.[147] "Intrusion on seclusion" is equally unsuitable as it presupposes an "intentional invasion into the victim's private affairs in a manner offensive to a reasonable person".[148] However, most of the time, deep fakes do not entail intrusions into a private matter where individuals have a reasonable expectation of privacy.

In addition to these remedies, the person affected by the deep fake may start a lawsuit for defamation or the closely related tort of "false light". These two legal tools seem to offer more effective protection in the context of deepfakes than the actions analyzed above. Defamation protects against false statements of a third party, harmful to the victim, such as those that might be contained in a deepfake. If the plaintiff is a public figure, the false statement must be "intentional or made with reckless disregard" for the victim's rights. Otherwise, it will be sufficient to show the negligence of the deepfake creator. The only limit to a successful defamation claim could be the disclosure of the deepfake as false content. At that point, in fact, the deepfake would not contain any false statement as it is clear that it is a synthetic content. This limitation also applies if a certain individual is put in a "false light" in front of the public, with a reckless disregard for the truth and this creates harmful implications.[149]

It is now clear that the current civil legal framework is ill-equipped to address the misuse of deepfakes. Furthermore, it is necessary to take into account that lawsuits are expensive and time-consuming and, as a result, victims might be unwilling to undertake private actions. In this context, the efforts of the U.S. federal and state lawmakers examined in paragraph A, are crucial to ensure some sort of protection for the individual and society as a whole. Moreover, some of these laws are criminal in nature and therefore have a strong deterrent effect against illegal conduct perpetrated through deepfake technology.

2. European Union legal framework

While the United States is seeking to deal with deepfakes both at the state and federal level, Europe currently has no laws specifically addressing this phenomenon. European bodies simply recognize deepfakes as artificial intelligence tools used to disseminate false content and increase disinformation.[150] Consequently, some of the policies adopted by the EU institutions to tackle the spread of disinformation on a large scale will be discussed below. This analysis will show that the

[147] Danielle Keats Citron, Sexual Privacy 128 YALE L.J (2019) at 1933-35.
[148] Ibid.
[149] see supra note 146 § 652E.
[150] National Endowment for Democracy (NED), "The Big Question: how will 'deepfakes' and emerging technology transform disinformation" (October, 2018) https://www.ned.org/the-big-question-how-will-deepfakes-and-emerging-technology-transform-disinformation/ accessed 23 July 2020.

EU has so far failed to prevent the harmful consequences of disinformation and therefore is not ready to fight deepfakes as part of the global fake news crisis.
Additionally, some national laws of Member States will be discussed as well as they could provide some protection, albeit weak.

a. Effort of European Commission against disinformation

Firstly, the EU Commission's Communication on Tackling online disinformation sets out some measures that could be relevant to address the challenges posed by deepfake technology.[151] For instance, in 2018, the European Union decided to collaborate with some private companies and launched the "Code of Practice on Disinformation" for some online platforms.[152] The Code relevant signatories recognize 'the importance of ensuring that online services include and promote safeguards against disinformation'.[153] In doing so, they have committed to report their actions to assess and mitigate the impact of disinformation. Beyond that, these parties should have provided some solutions to deal with the development of disruptive technologies, such as deep fakes. Running for a twelve-month trial period, this Code was an experiment in voluntary self-regulation by the tech industry and has produced not always satisfactory results.[154] In fact, few of the stakeholders involved seem to have fully implemented the code and achieved substantial progress. Despite some significant positive results (such as the fact-checking techniques implemented by Twitter and Facebook), there has been a lack of strong collaboration between the tech industry, governments, academia and civil society.[155] In addition to the Code, the European Commission has promoted the "Action Plan against Disinformation".[156] This policy initiative has classified disinformation as a hybrid threat and set four key objectives: to improve the EU's ability to detect and analyze disinformation; to strengthen cooperation and joint responses to this threat; to promote greater collaboration with online platforms and the tech industry; and to raise awareness and improve society's resilience.[157]

---

[151] This Communication highlights the Commission's awareness of deepfakes as powerful tools to manipulate public opinion. EU Commission "Communication on Tackling online disinformation: a European approach" (April 26, 2018), at 5 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236 accessed 23 July 2020.
[152] European Commission 'Code of Practice on Disinformation' (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation accessed 23 July 2020.
[153] Ibid at Ch II.C "Whereas clause".
[154] European Commission "Code of Practice on Disinformation One Year On: Online Platforms Submit Self-Assessment Reports" (2019).
[155] James Pamment 'The EU Code of Practice on Disinformation: Briefing Note for the New EU Commission' (2020) Carnegie Endowment for International Peace.
[156] Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Action Plan against Disinformation (JOIN(2018) 36 final).
[157] Ibid.

Despite the Union's actions in this field, it can be argued that there is much more to be done to effectively protect the rights of European citizens against these threats. This urgent need for intervention has emerged strongly in recent months. The Covid-19 pandemic has exacerbated the dilemmas of disinformation and shown the inadequacy of current European policies. According to senior World Health Organization officials, the spread of Covid-19 infectious disease would not only be the first pandemic case of the decade, but also the first case of "infodemic" in human history.[158] At the beginning of June 2020, the EUvsDisinfo web platform detected around 570 cases of fake news related to coronavirus. In these circumstances, the European Union must necessarily take further and more severe measures, in addition to improving its ability to detect and manage new forms of disinformation such as deepfakes.[159] Notably, the European institutions should establish a new regulatory framework in collaboration with platforms aimed at tackling the social media manipulation and ensuring the reliability of information sources. This strategy should be implemented through a clear definition of the duties and responsibilities of all stakeholders involved.

b. From the bottom up: drawing inspiration from Some Member State' national laws

Similarly to the United States, some of the existing laws in the EU Member States could help in certain limited and specific cases. For example, national laws on defamation can be a useful form of legal recourse when a deepfake has been created to damage the reputation of an individual (or a company). However, an action for defamation may fail if the deepfake is labeled as false because there is no longer a claim to disclose a misrepresentation.

Furthermore, copyright laws also need to be taken into account: deepfakes are often generated by using numerous images of a person that could be protected by copyright. The copyright owner has moral rights to that image such as, for example, in France, the right to respect the integrity of the work ("*droit au rispettare de l'intégrité di L'œuvre")*[160] or, in Germany, the right to prohibit alterations or other damages to the work that could harm his or her well-founded intellectual or moral interests.[161] However, as stated above, only the copyright owner (i.e. the author of the image) can bring an action for copyright infringement or moral rights breach and, most of the time, this individual

[158]John Zarocostas, 'How to fight an Infodemic' (2020) The Lancet at 676 https://doi.org/10.1016/S0140-6736(20)30461-X accessed 25 July 2020.

[159] On June 2020, "the EU Commission released a joint communication 'Tackling COVID-19 disinformation - Getting the facts right' to propose concrete actions to increase the resilience of the EU against the disinformation challenge. These include stepping up EU support to fact-checkers and researchers, strengthening of the EU's strategic communications capacities and enhancing cooperation with international partners, while ensuring freedom of expression and plurality." https://www.consilium.europa.eu/en/policies/coronavirus/fighting-disinformation/.

[160] Code de la propriété intellectuelle Article L121-1 (1992) Journal officiel de la République française.

[161] Act on Copyright and Related Rights (Urheberrechtsgesetz – UrhG) § 14 (Federal Law Gazette, 1965).

does not correspond to the person targeted by the deepfake. This is the main obstacle to such legal proceedings.

Personality rights could be a more effective remedy than copyright as they provide protection for the right to one's own image, the right to self-expression and the right to sexual self-determination which can be deeply affected by deepfakes. The person whose personality rights have been violated through deepfakes has the right to demand immediate cessation of the activity, removal of that manipulated content and compensation for damages. However, the success of an action depends very much on the specific facts of the dispute, even in countries such as France and Germany, which enforce personality rights. For this reason, these legal measures could provide protection to one individual and not to another, as they could have different effects depending on the case. Therefore, on the basis of these existing legal frameworks in France and Germany, it is difficult to counter the fraudulent use of deepfakes in general. Many laws would need to be amended to directly include a form of civil liability affecting harmful deepfakes.

Additionally, it can be argued that criminal laws in Europe provide a stronger deterrent against deepfakes. For example, the French criminal code regulates the "offence against the image of persons" by providing for a penalty of one year's imprisonment and a fine of €15,000 for the publication of any montage made with the words or image of a person without the latter's consent, unless it is clear that it is a montage, or this fact is expressly indicated.[162] So the party creating the deepfake and publishing it may be liable to sanctions, unless it proves that it was obvious that the material was a montage or this had been correctly reported. German criminal law also prohibits the unauthorised distribution of videos or images, including montages, if these can cause considerable damage to the reputation of the person portrayed.[163]

Digital identity theft could be even a more interesting tool to address some of the deepfake issues since it punishes the impersonation of third parties or the use of their data to disturb public peace or damage their reputation. Deepfakes, therefore, can already be sanctioned under certain conditions. However, it would be appropriate to amend existing laws or create new ones to better address the various devastating implications that can arise from the spread of a malicious deepfake.

[162] Code Penal (1992) Article 226-8.
[163] German Criminal Code (Strafgesetzbuch – StGB) § 187.

CONCLUSIONS AND RECOMMENDATIONS

Considering the constant and rapid evolution of deepfakes, the discussion presented so far will certainly not be definitive or resolutive, but it aims to expose an issue that should not be underestimated at all. The American actions in this area are only the first steps necessary to regulate a technology of very worrying implications that is spreading rapidly and that is going to produce more and more realistic outcomes. To date, anti-deepfakes legislation is still piecemeal in the United State[164] and almost entirely absent in Europe. For this reason, a precise and coherent modernisation of these legal systems in the field of deepfakes is needed to effectively regulate their abuse. At the same time, however, legislators should carefully draft this legislation, safeguarding the positive uses of technology in the light of freedom of expression.

Moreover, the enforcement of the law alone may not be sufficient. Discussion of the possible harmful implications of deepfakes has revealed that the wide range of deepfakes can have a dramatic and devastating repercussions in a variety of contexts. For this reason, it could be argued that a multi-level defence approach is crucial to effectively combat deepfakes. States should in fact implement a strategy that involves all stakeholders, from social platforms to citizens. This action plan should inevitably include investments in technological sector, education and media literacy, in a high-quality and reliable media landscape, as well as strong collaboration between governments, academic world, private companies and individuals.

Several platforms within the social media industry already seem willing to provide solutions in terms of detection and authentication. For instance, Facebook has stated that it intends to use artificial intelligence to spot deepfakes. For this purpose, Mark Zuckerberg's company has released the largest ever data set of deepfakes—more than 100,000 clips produced using 3,426 actors and a range of existing face-swapping techniques. It also announced the winner of the 'Deepfake Detection Challenge' in which 35,000 models of deepfakes detection techniques were submitted and evaluated. The winner model, however, has limited detection capability with an accuracy of only 65%. For this reason, Facebook has declared that it wants to develop its own detection software. Beyond that, several social media companies, most recently Twitter, have changed their policies to ban deepfakes and other manipulated media that could cause "serious harm"—such as content that threatens people's physical safety or could cause "widespread civil unrest".[165] A further interesting potential solution

[164] However, efforts to address this threat are clear: the adoption of the Deepfakes Accountability Act and other federal laws to deal with this reality is desirable.

[165] Shirin Ghaffary 'Twitter is finally fighting back against deepfakes and other deceptive media' Vox (2020).

came from some computer engineers who proposed to use blockchain technology to tackle deepfakes.[166] Such technology could guarantee the authenticity and traceability of videos and photos, thus increasing the transparency and security of the digital ecosystem. It is clear that science and legislation cannot be separated in this field.

Further effective means to prevent digital deception is media literacy. With proper training, individuals should recognize the risks and threats posed by deepfakes and learn to critically examine online content in general. This education could help individuals develop the ability to distinguish between truth and falsehood, enabling them to assess the accuracy, credibility and relevance of online information. Establishing a new awareness among citizens is therefore crucial to limit the harmful effects of false and manipulated content.

As this paper has shown, the problem of deepfakes does not have straightforward and immediate solutions. The law is certainly a fundamental remedy but it cannot be the only one. A multi-stakeholder effort is required in order to achieve concrete results and establish a safer digital environment.

[166] Paula Fraga-Lamas, Tiago M. Fernandez-Caramès 'Fake News, Disinformaton, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality' (2020) IT Professional Volume: 22 DOI:10.1109/MITP.2020.2977589.

Bibliography

Table of Cases

United States

- *Abrams v. United States*, 250 U.S. 616, 630 (1919)
- *Bigelow v. Virginia*, 421 U.S. 809 (1975)
- *Board of Education v. Pico*, 457 U.S. 853, 866-67 (1982)
- *Campbell v. Acuff-Rose Music Inc.*, 510 U.S. (1994)
- *Consolidated Edison Co. of New York v. Public Service Commission*, 447 U.S. 530 (1980)
- *Garrison v. Louisiana*, 379 U.S. 64, 75 (1964)
- *Gertz v. Robert Welch*, 418 U.S. (1974)
- *Hustler Magazine and Larry C. Flynt, Petitioners v. Jerry Falwell*, 485 U.S. (1988)
- *Hustler Magazine, Inc. v. Falwell*, 485 U.S. 46 (1988)
- *Linmark Associates Inc. v Townships of Willingboro*, 431 U.S. 85 (1977)
- *Lorillard Tobacco Co. v. Reilly*, 533 U.S. 525 (2001)
- *McCreary County v. American Civil Liberties Union*, 545 U.S. 844 (2005)
- *N.Y. Times Co. v. Sullivan*, 376 U.S. (1964)
- *New York Times v. Sullivan*, 314 U.S. 252, 270 (1964)
- *Philadelphia Newspapers Inc v. Hepps*, 475 U.S (1986)
- *Police Department v. Mosley*, 408 U.S. (1972)
- *Randall v. Sorrell*, 548 U.S. 230 (2006)
- *Red Lion Broadcasting Co. v. FFC*, 395 U.S. 367, 390 (1969)
- *Reno v American Civil Liberties Union (ACLU)*, 521 U.S. 844 (1997)
- *Reno v. American Civil Liberties Union*, 521 U.S. 844 (1997)
- *Texas v. Johnson*, 491 U.S. 397, 419-20 (1989)

- *United States v. Alvarez*, 567 U.S. 709 (2012)
- *Virginia State Pharmacy Board v. Virginia Citizens Consumer Council* 425 U.S. 748 (1976)
- *Virginia v. Black*, 538 U.S. 343, 358 (2003)
- *Virginia v. Hicks*, 521 U.S. 844, 855 (1997)
- *Walker v Sons of Confederate Veterans*, 115 S.Ct. 2239 (2015)

European Union

- *Aquilina and Other v. Malta*, App. No 28040/08, Eur. Ct H.R. (2011)
- *Chamber judgment Steel and Morris v. United Kingdom*, App. No 68416/01, Eur. Ct. H.R. (2005)
- *Delfi AS V. Estonia*, App. No 64569/09, Eur. Ct H.R. (2015)
- *Handyside v. the United Kingdom*, App. No. 5493/72 Eur. Ct H.R. (1976)
- *Hertel v Switzerland*, App. No 25181/94, Eur. Ct. H.R. (1998)
- *Mouvement Raëlien Suisse V. Switzerland*, App. No 16354/06, Eur. Ct. H.R. (2012)
- *Steel and Morris v. UK*, App. No 68416/01, Eur. Ct. H.R. (2005)
- *Tammer v. Estonia*, App. No 41205/98, Eur. Ct. H.R. (2001)
- *Társaság a Szabadságjogokért v Hungary*, App. No 37374/05, Eur. Ct. H.R. (2009)
- *Timpul Info-Magazine and Anghel v Moldova* App. no. 42864/05 Eur. Ct. H.R. (2007)

Table of Legislation

United States

- Communication Decency Act of 1996 (CDA)
- H.R. 3230, 116th Congress – Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019
- H.R. 4355, 116th Congress – Identifying Outputs of Generative Adversarial Networks Act (2019)

- H.R.6088, 116th Congress – Deepfakes in Federal Elections Prohibition Act (2020)
- Restatement (Second) of Torts (1977) AM. LAW INST.
- S. 2065, 116th Congress – Deepfake Report Act of 2019 (2019)
- S.1790; 116th Congress – National Defense Authorization Act for Fiscal Year 2020 (2019-2020)
- The Digital Millennium Copyright Act (DMCA), 17 USC (1998)
- U.S. Const. Amend. I.

California

- AB-602 – 'Depiction of individual using digital or electronic technology: sexually explicit material: cause of action' (2019)
- AB-730 – 'Elections: deceptive audio or visual media' (2019)

New York

- AB A5605C – 'Establishes the right of publicity and provides for a private right of action for unlawful dissemination or publication of a sexually explicit depiction of an individual' (2019)

Texas

- Texas Senate Bill 751 – 'Relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election' (2019)

Virginia

- Code of Virginia (1950)
- House Bill No. 2678 – 'Unlawful dissemination or sale of images of another'

European Union

- Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling. Tackling online disinformation: a European Approach' COM (2018) 236 final
- Commission, Code of Practice on Disinformation (2018) <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454> accessed 23 July 2020

- European Council; ‘Fighting disinformation’ (2020) https://www.consilium.europa.eu/en/policies/coronavirus/fighting-disinformation/ accessed 23 July 2020
- Charter of Fundamental Rights of the European Union (2012) OJ C 326/391
- European Convention on Human Rights (1953)
- High Representative of the Union for Foreign Affairs and Security Policy, ‘Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Action Plan against Disinformation’ (JOIN (2018) 36 final)

France

- Code de la propriété intellectuelle (*Journal officiel* de la République française, 1992)
- Code Penal (1992)

Germany

- Act on Copyright and Related Rights (Urheberrechtsgesetz – UrhG) (Federal Law Gazette, 1965)
- German criminal code (Strafgesetzbuch – StGB)

Other sources

- Ajder H et al, ‘The State of Deepfakes: Landscape, Threats, and Impact’ (2019) Deeptrace
- Baccarella C et al., ‘Social media? It’s serious! Understanding the dark side of social media’ (2018) European Management Journal
- Baker C E, ‘Human Liberty and Freedom of Speech’ (1989) Oxford University Press
- Beavers O, 'Washington fears new threat from 'deepfake' videos' (2019) The Hill <https://thehill.com/policy/national-security/426148-washington-fears-new-threat-from-deepfake-videos> accessed 1 July 2020
- Brietzke P H, ‘How and Why the Marketplace of Ideas Fails’ (1997) Valparaiso University Law Review
- Caldera E, 'Reject the Evidence of Your Eyes and Ears: Deepfakes and the Law of Virtual Replicants' (2019) 50 Seton Hall Law Review

- Chesney B and Citron D, ‘Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security’ (2019) 107 California Law Review
- Choudry S, ‘The Migration of Constitutional Ideas’ (2007) Cambridge University Press
- Cinecom, ‘DEEPFAKE Tutorial: A beginners Guide’ (2019) <https://www.youtube.com/watch?v=t59gRbpYMiY> accessed 15 June 2020
- Citron D. K., ‘Sexual Privacy’ (2019)128 Yale Law Journal
- – –, ‘*Hate Crimes in Cyberspace*’ (2014) Harvard University Press
- Coldewey D, ‘DEEPFAKES Accountability Act would impose unenforceable rules - but it’s a start’ (2019) TechCrunch <https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/> accessed 25 July 2020
- Courtney W and Paul C, ‘Firehose of Falsehoods:Russian propaganda is pervasive, and America is behind the power curve in countering It’ (2016) U.S. NEWS & WORLD REP <https://www.usnews.com/opinion/articles/2016-09-09/putins-propaganda-network-is-vast-and-us-needs-new-tools-to-counter-it> accessed 5 June 2020
- Donavan J, ‘Deepfake Videos Are Getting Scary Good’ (2018) Howstuffworks https://electronics.how stuffworks.com/future-tech/deepfake-videos-scary-good.htm> accessed 18 June 2020
- Drinnon C; ‘When Fame Takes Away the Right to Privacy in One's Body: Revenge Porn and Tort Remedies for Public Figures’ (2017) 25 William & Mary Journal of Women and the Law
- European Commission, ‘Code of Practice on Disinformation One Year On: Online Platforms Submit Self-Assessment Reports’ (2019) <https://ec.europa.eu/commission/presscorner/detail/en/statement_19_6166> accessed 23 July 2020
- Ferraro F M et al, ‘The Federal “Deepfakes" Law’ (2020) 3:4 Journal of Robotics, Artificial Intelligence & Law
- Fischer W, ‘California's governor signed new deepfake laws for politics and porn, but experts say they threaten free speech’ (2019) Business Insider <https://www.businessinsider.com/california-deepfake-laws-politics-porn-free-speech-privacy-experts-2019-10?r=US&IR=T> accessed 20 July 2020

- Fraga-Lamas P et al, 'Fake News, Disinformaton, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality' (2020) 22 IT Professional
- Ghaffary S, 'Twitter is finally fighting back against deepfakes and other deceptive media' Vox (2020) <https://www.vox.com/recode/2020/2/4/21122653/twitter-policy-deepfakes-nancy-pelosi-biden-trump> accessed
- Goodfellow I J et al, *'Generative Adversarial Networks'* (arXiv, 10 Jun 2014) <https://arxiv.org/abs/1406.2661> accessed 15 June 2020
- Green R, *'*Counterfeit Campaign Speech' (Faculty Publications, 8 Jul 2019) <https://scholarship.law.wm.edu/facpubs/1923> accessed 18 June 2020
- Greene D, *'*We Don't Need New Laws for Faked Videos, We Already Have Them*'* (EFF, 13 Feb 2018) < https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them> accessed 23 July 2020.
- Hall H K; 'Deepfake Videos: When Seeing Isn't Believing' (2018) 27 Cath U J L & Tech 51.
- Harr J S et al, JKingsbury J; *'Constitutional Law and the Criminal Justice System'* (7th edition, Cengage Learning).
- Harris D, 'Deepfakes: False Pornography Is Here and the Law Cannot Protect You' (2018-2019) 17 Duke L & Tech Rev 99.
- Hudson Jr. D L, *'*Counterspeech Doctrine*'* (Middle Tennessee State University, Dec 2017) <https://www.mtsu.edu/first-amendment/article/940/counterspeech-doctrine> accessed 1 July 2020.
- Kietzmann, J et al. *'*Deepfakes: Trick or treat*?'* (2020) Business Horizons.
- Knight W; *'*The US Military is Funding an Effort to Catch Deepfakes and Other AI Trickery*'* (2018) MIT Technology Review.
- Kulik Tom; *'*Faking It: Why Deepfakes Pose Specific Challenges Under Copyright & Privacy Laws' (2019) Above the Law <https://abovethelaw.com/2019/07/faking-it-why-deepfakes-pose-specific-challenges-under-copyright-privacy-laws/?rf=1> accessed 25 July 2020
- Lombardi C; *'*The Illusion of the Marketplace of Ideas*'* (2018) KIMEP, School of Law.

- Matsuda M J; *'*Words that wound: critical race theory, assaultive speech, and the First Amendment*'* (1993).
- McCarthy J. Thomas et al, 'The Rights of Publicity and Privacy*'*, Clark Boardman Callaghan (2d Ed., 2020).
- Melville K; *'Humiliated, Frightened and Paranoid: The Insidious Rise of Deep Fake Porn Videos'* (ABC News, 30 Aug 2019) <www.abc.net. au/news/2019-08-30/deepfake-revenge-porn-noelle-martin-story-of-im age-based-abuse/11437774>
- Mill J S, '*On Liberty*' (2d ed 1863)
- Milton J; *'Areopagitica'* (1644; Jebb ed. Cambridge University Press, 1918).
- Mostert F et al, *'*How to counter deepfakery in the eye of the digital deceiver*'*(2020) Financial Times <https://www.ft.com/content/ea85476e-a665-11ea-92e2-cbd9b7e28ee6> accessed on 18 June 2020.
- Napoli P M, *'*What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble*'* (2018) 70 Federal Communications Law Journal.
- National Endowment for Democracy, "The Big Question: how will 'deepfakes' and emerging technology transform disinformation" (2018) NED <https://www.ned.org/the-big-question-how-will-deepfakes-and-emerging-technology-transform-disinformation/> accessed 23 July 2020
- Nonnecke B M; 'Opinion: California's Anti-Deepfake Law Is Far Too Feeble*'* (2019) Wired <https://www.wired.com/story/opinion-californias-anti-deepfake-law-is-far-too-feeble/> accessed 23 July 2020.
- Pamment J, 'The EU Code of Practice on Disinformation: Briefing Note for the New EU Commission*'* (2020) Carnegie Endowment for International Peace.
- Pariser E, *'The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think'* (2012) Penguin Books.
- Pollicino O, 'Fake News, Internet and Metaphors (to Be Handled Carefully)' (2017) Italian Journal of Public Law 1.
- Richards et al, 'Counterspeech 2000: A New Look at the Old Remedy for "Bad"Speech*'* (2000) B.Y.U. Law Review.

- Rothkopf J , 'Deepfake Technology Enters the Documentary World' (July 1, 2020) New York Times.
- Schapiro Z, 'DEEP FAKES Accountability Act: Overbroad and Ineffective' (2020) Intellectual Property & Technology Forum.
- Schauer F; 'The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience' (2004) Harvard Law Review.
- Shao G, 'What 'deepfakes' are and how they may be dangerous'(2019) <https://www.cnbc.com/2019/10/14/what-is-deepfake-and-how-it-might-be-dangerous.html> accessed 15 June 2020.
- Snow J, 'Deepfakes for good: Why researchers are using AI to fake health data' (2018) FAST COMPANY, <https://www.fastcompany.com/90240746/deepfakes-for-good-why-researchers-are-using-ai-for-synthetic-health-data> accessed 20 June 2020.
- Stupp C, *'Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case'* (2019) Wall Street Journal <https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402> accessed 19 June 2020.
- Toto C; *'Protecting Elections: Regulating Deepfakes in Politics'* **(**2020) JDSupra < https://www.jdsupra.com/legalnews/protecting-elections-regulating-39567/> accessed 8 August 2020.
- Tribe L H; *'American Constitutional Law'* (2d ed. 1988) New York: Foundation Press.
- Tsesis A; 'Categorizing Student Speech' (2018) 102 Minnesota Law Review.
- Voorhoof D; 'The Right to Freedom of Expression and Information under the European Human Rights System: Towards a more Transparent Democratic Society' (2014) EUI Working Paper RSCAS.
- Wakefield J et al, 'Could deepfakes be used to train office workers?' (February 29, 2020) BBC News <https://www.bbc.com/news/technology-51064933> accessed 20 June 2020.
- Westerlund M, 'The Emergence of Deepfake Technology: A Review' (November, 2019) Tech. Inn. Management Review.
- Zarocostas J, 'How to fight an Infodemic' (2020) The Lancet <https://doi.org/10.1016/S0140-6736(20)30461-X> accessed 25 July 2020.

- Zucconi A; 'Understanding the Technology Behind DeepFakes' (2018) Alan Zucconi Blog <https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/> accessed 15 June 2020.