Agata Wróbel

Can GDPR and Copyright law stop *deepfake*? A comprehensive legal analysis of the *deepfake* phenomenon and liability for the *deepfake* in the polish legal system

Abstract

The article is a comprehensive analysis of the relationship between deepfake technology and identity issues in the context of the Polish legal system. In the context of the dynamic development of generative Artificial Intelligence systems, an important aspect is the protection of individuals' identities. This article focuses on two key areas: detection technology and the legal framework that balances freedom of speech with the protection of identity and information security. The first part of the article focuses on the role of detection technology in combating deepfake. Current trends and advances in the field of content manipulation are analyzed. The second important aspect addressed in the article is the analysis of the legal framework regulating deepfake in Poland. Criminal and civil liability issues related to the creation and dissemination of deepfake are discussed, with a particular focus on image protection and related copyright law an related rights. The relationship between creative freedom and the protection of individual identity is also analyzed, including through the lens of Polish case law. The last part of the article focuses on current developments in the European AI Act (status as of December 2023). It is indicated how the new AI regulations affect the issues of identity protection and combating deepfake.

Keywords: Deepfake, Identity protection, Artificial intelligence, Machine Learning, AI Act, Image Rights.

1. Introduction

In today's era of rapid technological advancement, the concept of *deepfake* is becoming more and more common and disturbing at the same time. This technology, based on artificial intelligence and *machine learning*, makes it possible to create fake video and audio content that can easily fool viewers into appearing authentic. People in public spaces provide a lot of content that can be used to create such image or sound manipulation. Many such creations have already

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

been circulating around the world since 2018, such as the video *This is not Morgan Freeman - A Deepfake Singularity* by Bob de Jong, a Dutch filmmaker specializing in creating virtual images, with Morgan Freeman's voice dubbed here by dubbing actor Boet Schouwink[1] , a video with the clickbait title *You Won't Believe What Obama Says in This Video!* by Jared Sosa and Jonah Peretti[2] or even a video of the dancing Queen of the United Kingdom, published as a warning against deepfakes on the UK's Channel 4 under the title *Deepfake Queen: 2020 Alternative Christmas Message*[3].

Thanks to these develops, efforts to identify and combat deepfakes, which pose a serious threat to the reliability of information and privacy of individuals, have been intensifying in recent years. Technological development is mainly focused on improving AI-based detection algorithms. Increasingly complex *machine learning models*, especially those based on *deep neural networks*, are trained to analyze the subtle characteristics of fake media content. Identifying abnormalities in facial movements, video artifacts, or unusual sound patterns has become a key area of research. Research projects, such as fake content detection competitions organised by universities and various institutions dealing with the subject of cyber security, stimulate the development of innovative solutions. At the same time, tech companies are investing in the development of dedicated tools to detect and limit the spread of fake content on their platforms. Social media platforms are also taking steps to combat this phenomenon. They are implementing tools to report and remove false content and tighten policies against disinformation. They also work with research and government institutions to respond effectively to the growing threat. Despite the progress made, there are still many challenges in detecting them. Creators of fake content are constantly improving their techniques, which requires constant adaptation of detection tools.

It may seem that these are innocent jokes similar to memes, which no one treats as a serious threat. Is this correct? I think *deepfakes* can be called the next, more dangerous stage. It is worth considering the legality of such practices and the limits of lawlessness, after all, there is freedom of speech and artistic creation as a counterbalance. Laws against fake media content are enacted in many jurisdictions, and those responsible for creating and distributing it may face legal consequences. I believe that this topic is still something new in Polish legal science and should

[1] DiepNep, This is not Morgan Freeman - A Deepfake Singularity, 2021. Accessed 29.12.2023, https://www.youtube.com/watch?v=oxXpB9pSETo&t=1s.
[2] BuzzFeedVideo, You Won't Believe What Obama Says In This Video!, 2018. Accessed 29.12.2023, https://www.youtube.com/watch?v=cQ54GDm1eL0.
[3] Channel 4 Entertainment, Deepfake Queen: 2020 Alternative Christmas Message, 2020. Accessed 29.12.2023, https://www.youtube.com/watch?v=IvY-Abd2FfM.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

not be trivialized. Legal and regulatory aspects are becoming increasingly important in the context of the fight against deepfakes.

2. How does *deepfake* work and how many ways can someone steal your identity?

The name *deepfake* comes from the combination of the terms *deep learning* and *fake.* This phenomenon is sometimes written with two words, *deep fake*[4] , which is particularly visible in Polish studies, which results from the specificity of the adaptation of Anglicisms to the Polish language. However, the most commonly used is the connecting spelling as a *"deepfake".* Analyzing the components of this word, *deep learning* is a category *of machine learning*. This means that the design of the network is not aimed at specific feature recognition but is based on their identification grounded on precisely labeled data sets provided for analysis. The program is capable to detect features on its own, so it becomes possible to process huge amounts of data, and the network can automatically acquire higher-level representations of features. That means it can identify complex patterns in the input and, let's say, can "learn by itself"[5] .

In the case of generative AI which is used for *deepfake* not only text data is generated. generative AI not only learns from data, but also creates new data instances that mimic the properties of the input data. The midpoint of this revolutionary technology are *generative adversarial networks*, known as GANs. These advanced neural network architectures consist of two key components: a *generator*, responsible for generating fake data, and a *discriminator*, which evaluates the authenticity of this data[6]. The most common form of *deepfake* is a form of realistic face and voice substitution, which allows for the creation of video or audio content in which a person appears to speak or perform actions that they have never actually performed. The process of creating *a deepfake* involves intensively training *machine learning algorithms* on a huge amount of source data, such as videos, photos, or sounds, containing information about a specific person or character. Then, with the help of these algorithms, it becomes possible

[4] K. Szpyt, *Sztuczna inteligencja i nowe technologie (nie zawsze) w służbie ludzkości, czyli cywilnoprawna problematyka rozwoju i popularyzacji technologii deepfake*, in: K. Flaga-Gieruszyńska (ed.), *Sztuczna inteligencja, blockchain, cyberbezpieczeństwo oraz dane osobowe. Zagadnienia wybrane*, Warsaw 2019, pp. 75-94.

[5] J. Wilpon, T. Thomson, S. Bangalore, P. Haffner, M. Johnston, Interactions LLC, *Fundamentals of Machine Learning, whitepaper*. Accessed: 20.09.2024, https://www.interactions.com/resources/technology/fundamentals-machine-learning/, pp. 8-10.

[6] X. Mingchuan Long, M. Zha, *An Overview of Generative Adversarial Networks*, Journal of Computing and Electronic Information Management, 10(3), 2023, pp. 31-36.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

to generate fake content in which the face and/or voice of the source person is perfectly imitated by another person[7].

In the world of deepfakes, there are two main techniques: *face reconstruction*[8] and *face swapping*[9]. The former allows to transfer one person's facial expressions to another's face, while the second technique involves digitally replacing one person's face with another person's face in a video or photograph. Innovative methods like that make the line between reality and deception more and more blurred. Visual modifications can involve manipulating a photo or video. Video modification is very often correlated with sound modification. It is worth being aware that it is becoming more and more effortless to manipulate the sound itself, i.e. *voice cloning*. In the free online access one can find a large number of programs in which an audio can be create by providing several voice samples. Then having the voice sample one can type any sentence which will be read by that voice. Such samples are very easy to obtain from people active in public space, i.e. from recordings of public speeches at conferences, advertising spots, etc.[10].

3. Legal liability

At this point, it is necessary to determine the protected interest that is threatened by deepfake content. *The image* remains under the protection of civil law, regardless of the protection provided for in other provisions, because it is included in the catalogue of personal interests under art. 23 of the Civil Code[11]. *The image* is also regulated under art. 81 of the Copyright Law[12], according to which the dissemination of an image requires the consent of the person

[7] For more on how to create a deepfake see: M.L. Joshi, *Deepfake: A Systematic Review*, Kalyan Bharati, No. 0976-0822, 2022, UGC-CARE List Group I (73-74), pp. 73-74.

[8] To visually illustrate this method, it can be compared to creating computer games. We can find a lot of behind-the-scenes from the production of games, such as the game "The Last of Us", where facial expressions and actors' movements from motion capture sessions were later used by the team to animate the characters in cutscenes. See D. Shelton, *Capturing The Last of Us: Motion Capture Pipeline*, 2015. Accessed 29.12.2023, https://www.gdcvault.com/play/1021854/Capturing-The-Last-of-Us or PlayStation Australia, The Last of Us Part I. Motion Capture Trailer, 2022. Accessed 29.12.2023, https://www.youtube.com/watch?v=4Wd0P5WnO9Q.

[9] To illustrate this method, it is worth citing the example of Nicolas Cage, who for unexplained reasons became the main target of remakes. The actor's face can be seen in many films made as a hobby by Internet users, such as in the movie “Man of Steel”. See: S. Haysom, *People are using face-swapping tech to add Nicolas Cage to random movies and what is 2018*, 2018. Accessed 29.12.2023, https://mashable.com/article/nicolas-cage-face-swapping-deepfakes.

[10] The quality of the cloned voice has been subjected to many studies. One of them is reported by Bochyńska N. in the article on *CyberDefence24.pl.* The researchers trained their system on speech samples from 107 speakers to investigate why they sound ‘human’. Testing less secure systems, they achieved a 99% success rate and a 72% success rate in more resilient ones.

[11] Act of 23 April 1964 Civil Code (Journal of Laws. 1964 No. 16 item 93).

[12] Act of 4 February 1994 on Copyright and Related Rights (Journal of Laws. 1994 No. 24 item 83).

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe” (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

shown in it. Although there is no legal definition of *the image*, such a definition can be derived from the judicature: an image as a legal good protected under art. 24 of the Civil Code and art. 81 of the Copyright Law is a broadly understood likeness of a person, i.e. the concretization and determination of the physical image of an individual, that is suitable for replication and dissemination. Therefore, *the image* in the legal sense is not the same as the physical appearance of a person, it is a representation and concretization of this appearance. *The image* is a certain immaterial good, the protection of which granted by the provisions of the Copyright Law is strictly formal and objective in nature"[13]).

3.1. Civil liability and liability under the Law of Copyright and Related Rights

On the civil law plane, a number of general claims can be brought in this case, such as a claim for damages or cessation of infringing personal property. The right to image, as one of the personal interests, plays a key role in the analysis of the situation of people presented in *deepfake* recordings. Despite certain legal regulations in Poland, in particular in art. 81 and 83 of the Copyright Law and art. 23, 24, 448 of the Civil Code, the lack of a clear definition of image creates some difficulties. One definition is proposed by A. Matlak: an individual can be perceived through the prism of multiple images, i.e. perceptible physical features that make up his physical appearance at different times of life, from infancy to old age[14].

In the case of *deepfake* recordings it should be remembered that the technique of making the image does not significantly affect its qualification for copyright. Even in cases where the facial image is computer-generated rather than recorded directly, there is still the possibility of image infringement. Similar rules apply to caricatures or avatars in computer games[15] . The Copyright Law regulates the use and modification of pieces of works in the context of *deepfakes*. Violation of this law results in liability for damages and criminal liability. Dissemination of an image without the consent of the person depicted in it is an infringement of moral rights, which opens the way to claims for damages or redress. As it has already been noted, copyright law requires obtaining permission to disseminate an image (art. 81(1) of the Copyright Act). On the other hand, permission is not required for the dissemination of the image of the public person, if the image was made in connection with the performance of public

[13] Judgment of the Court of Appeal in Warsaw of 19 September 2018, ref. VI ACa 528/17, LEX nr 2616080.
[14] A. Matlak, *Cywilnoprawna ochrona wizerunku*, Kwartalnik Prawa Prywatnego 13(2), 2004, pp. 317–358.
[15] M. Modrzejewska, *Podobieństwo awatarów do znanych osób żyjących* [in:] *Verba volant, scripta manet. Księga jubileuszowa dedykowana Pani Profesor Bogusławie Gneli*, ed. A. Kaźmierczyk, K. Michałowska, M. Szaraniec, Warsaw 2023, pp. 721-731).

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

functions, in particular political, social, or professional, as well as the dissemination of the image of a person constituting only a detail of a whole, such as a gathering, landscape, public event (art. 81(2) of the Copyright Act). The Civil Code regulates liability for infringement of personal rights, including the right of publicity. Such a breach may also be the basis for damages claims. In addition, the unfair competition rules (Competition and Consumer Protection Act 2007) can be used to spread false information via deepfakes[16].

The problem with such alterations of works is very often the issue of lack of consent. In Poland, according to art. 81(1) of the Copyright Act, the creator of *a deepfake* recording should obtain the consent of the person shown in it. However, there are cases, especially with the use of widely recognized audiovisual works, in which it may be necessary to obtain the consent of various people who may be recognizable in the material, because the face belongs to person A and the rest of the body or features belong to person B, it is worth looking at the possibility of identifying people not only directly, but also indirectly. The use of widely recognized audiovisual work to create *a deepfake* can lead to a situation where a person can be recognized based on information about the work itself[17]). The following part of the article will address the issue of criminal liability, but it must be admitted that liability for *deepfakes* in Poland is based mainly on the provisions of copyright law and civil law. It is necessary to consider the right to image as a personal right, as well as the provisions on consent to the use of the image. In the case of *a deepfake,* the infringement of copyright and personal property can lead to claims for damages and damages.

For a broader context, it is worth to mention exceptions. *Deepfakes* in the use of audiovisual works may violate copyright, especially if they are used for satirical purposes. While it is possible to invoke fair use under Article 29 with index 1 of the Act on Copyright and Related Rights, it is justified only in the case of parody (Article 29a), pastiche or caricature[18] . In addition, the creators of deepfakes must consider liability for potential infringements of personal rights, in particular the right of publicity. In the case of public characters, such as politicians or celebrities, it is possible to disseminate an image without consent, if it is related to their function, in accordance with Article 81(2)(1) of the Copyright

[16] See more: O. Bodanka, Image violation with the use of deepfake technology – a legal and practical analysis, Opolskie Studia Administracyjno-Prawne 20(1), 2022, (11–32), pp. 25-27.
[17] See: O. Bodanka, Image violation with the use of deepfake technology – a legal and practical analysis, Opolskie Studia Administracyjno-Prawne 20(1), 2022, (11–32), pp. 20-24.
[18] See more: J. Barta, M. Czajkowska-Dąbrowska, Z. Ćwiąkalski, R. Markiewicz, E. Traple, Prawo autorskie i prawa pokrewne Komentarz, Wolters Kluwer, 2011, pp. 194-195.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

Law. However, using an image in a way that violates honor or dignity can still lead to legal consequences.

3.2. Penal liability for deepfakes

It is not difficult to imagine negative ways in which *deepfakes* can be used. These are not only funny videos with an entertaining objective. What about using someone's image in a way that could damage that person's reputation, such as pasting someone's face into a pornographic video? Article 191a of the Penal Code[19]) provides that the penalty of imprisonment is imposed on anyone "who records the image of a naked person or a person in the course of a sexual act by using violence, unlawful threat or deception against him, or disseminates the image of a naked person or a person during a sexual act without the consent of that person (...)". However, in the case of manipulated pornographic material, there is a problem with the qualification of elements of such an act, because the person actually appearing in the film – the porn actor – is replaced by the image of another person who does not know about it[20]). The doctrine argues that in order to meet the constituent elements of the offence, the victim must be personally and physically present during a sexual act that has been recorded or disseminated without her consent. It cannot be maintained that a person whose only face has been altered by deepfake technology because that person is neither naked nor actually participating in the sexual act depicted[21]).

It may be appropriate to pursue liability under Article 190a § 2 of the Penal Code, which penalises identity theft in the form of using someone's image without their consent, whereby the person in the materials made available may be publicly identified, thus causing material or personal damage to them, and the situation of the aggrieved parties can certainly be equated to the latter through a sense of humiliation. However, the problem arises in the first part of paragraph 2, i.e. in the expression '*impersonation of another person'*, as this term is ambiguous, vague concept. To clarify, I will paraphrase one of A. Lach's linguistic definitions: in Polish, the verb *'to impersonate'* means '*to falsely impersonate someone, to admit to someone else's merits'*. It should be assumed that in the case of actions directed against a person, the *modus*

[19] Act of 6 June 1997 Criminal Code (Journal of Laws 1997 No. 88 item 553).
[20] M. Sewastianowicz, Deepfake - ofiara realistycznej przeróbki może mieć problem z dochodzeniem swoich praw, 2023. Accessed 29.12.2023, www.prawo.pl; V. G. Garcia, Deep fake – postęp technologiczny a prawo karne, Acta Iuridica Resoviensia, 44(126), 2024 (39–53), pp. 39–53.
[21] A. Ziobroń, Deepfake a prawo karne. Uwagi de lege lata i de lege ferenda dotyczące fałszywej pornografii, Studenckie Prace Prawnicze, Administratywistyczne i Ekonomiczne 37, 2021, (225–238), pp. 229-230.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

operandi of the perpetrator consists in creating a false belief that he is someone else by misleading, actively reinforcing the false belief. In fact, in most cases, the impersonation mark will not be implemented, but I think that such a situation is imaginable. I think that in the case of the so-called *revenge porn*[22] , the creator may be so determined that he may want to create the illusion that it is the victim who publishes videos with his own participation.

A defamation lawsuit can be an appropriate criminal response. The protection under art. 216 of the Penal Code is not limited to the spoken or written word, as unlawful conduct may also be transmitted by non-verbal means of expression or through the conduct of a person. There may be so many causal forms of an act subject to criminal sanction that it would be unreasonable, if not impossible, to define them so precisely in law. Therefore, the liability of the author of the *deepfake* for the insult may be taken into account. The dominant factor in the criminal assessment of a given criminal act is whether the behavior violates someone's personal dignity, honor, whether it is contemptuous in nature, or whether it does not express a lack of respect for another person, because in the Polish language the verb 'to *insult*' means '*to violate someone's dignity*'[23].

3.3. Personal data protection and the use of the image in deepfake

Another field of exploitation of the topic is also the analysis of personal data protection regulations. First, recital 18 of the preamble to the GDPR[24] states that the Regulation does not apply to the processing of personal data by an individual in the course of a purely personal or domestic activity, that is to say, unrelated to a professional or commercial activity. Personal or domestic activities may include, but are not limited to, correspondence and address storage, maintaining social ties, and online activities undertaken as part of such activities. From the regulation's perspective, the creation of deepfakes will be *the processing of* personal data. *Processing* under the Regulation means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval,

[22] More on this topic can be found in: R.A. Delfino, Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act, Fordham Law Review 88(3), 2019, (886–938), pp. 892-894 or S. Wieczorek, M. Kubiak, Ryzyka i szanse wynikające z rozwoju nowych technologii w branży mediowej na przykładzie zjawiska deep fake – analiza prawna, Monitor Prawniczy 21, 2019, (101–107), pp. 101.
[23] I. Zgoliński, Commentary on Article 216. In: V. Konarska-Wrzosek (ed.) Kodeks karny. Komentarz, (1146–1150), Warsaw 2021, pp. 1146-1150.
[24] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) Official Journal of the European Union, OJ L 119.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction (art. 4 §2 of the GDPR). Since the GDPR does not apply to the processing of personal data by a natural person (human) as part of personal activities, then if someone creates deepfakes for their own purposes, then the GDPR does not apply.

What about the consent of the person whose image is used in the deepfake? Does this person have to give consent? According to Article 6(1)(a) of the GDPR, processing is only lawful in cases and to the extent that one of the above conditions is met, and one of these conditions – applicable here – is that the data subject has consented to the processing of his or her personal data for one or more specific purposes. The basic question is – is the creation of deepfake content a *purely personal or domestic activity* and thus not subject to the GDPR? The Provincial Administrative Court in Warsaw ruled on the exclusion from recital 18 of the preamble to the following way: "this should not be interpreted as an example of an absolute exclusion, occurring in every situation and in the case of any non-commercial activity of the user on the Internet, but with a restriction to those factual situations in which the processing of data, in particular their sharing via the Internet, does not take place for the benefit of an unlimited group of recipients"[25]. A deepfake published in an open online forum will be available to an unlimited number of recipients and thus the exclusion cannot be used. This implies the need to obtain the consent of the person whose image is presented. However, it is difficult to imagine a situation in which a politician will agree to present his image in a ridiculing, often vulgar video, and it is such content that is the biggest problem.

Another thread is related to recital 27 of the preamble, according to which the regulation does not apply to personal data of deceased persons. Member States may adopt regulations on the processing of personal data of deceased persons – but the Polish legislator has not adopted such regulations. So how can we assess from a polish legal point of view the activity implemented, for example, in the seventh part of the popular film series "Fast and Furious", where a digital double of the popular actor Paul Walker was created, who managed to shoot only half of the scenes for the movie? The problem with the recognition of legal protection of the privacy of the deceased is that there is no entity that could exercise such a right for the deceased, due to the subjective nature of the right to privacy[26] . However, in the doctrine one can find positions justifying the admission of *post-mortem protection*. This position was

[25] Judgment of the Provincial Administrative Court in Warsaw, ref. II SA/Wa 895/20, Legalis no. 2658998
[26] K. Rainko, A. Sikora, *Wizerunek w świetle prawa cywilnego a ochrona danych osobowych* [in:] *Ochrona danych osobowych w prawie publicznym*, ed. M. Jędrzejczak, Warsaw 2021, pp. 161-166.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

adopted by m.in. J. Sieńczyło-Chlabicz on the basis of Article 83 of the Copyright Act, which refers to Article 78(1) of the Copyright Act (an action for the protection of moral rights): recognising the personal and economic nature of the right to an image, the following solution could be considered. Namely, the right to the image covering the personal interests of the portrayed person expires at the moment of his death[27]. However, it is possible to protect the right to the image of the deceased by the relatives on account of their personal rights, especially with the demand to respect the right to the cult of the memory of the deceased[28] . In this particular case of a popular action film, however, it would be difficult to conclude that continuing the work was a violation of the actor's good memory.

4. Is freedom of speech a ticket for creating deepfake content?

In the context *of deepfakes*, whose potential for manipulating video and audio content is a concern, the issue of legal liability is becoming an important element of analysis, especially in the light of the law on freedom of speech. Freedom of expression, considered a fundamental human right, is protected by the European Convention on Human Rights[29]) and, in particular, by art. 10 of that Convention. However, it is worth noting that while freedom of expression is undoubtedly a key pillar of a democratic society, its exercise must be subject to certain limitations, in line with the Convention. For *deepfakes*, where there is a potential risk that the technology could be used for harmful purposes, such as spreading misinformation or violating the privacy of individuals, certain restrictions may be necessary. The right to freedom of expression is not absolute and may be subject to restrictions for the public. For *deepfakes*, where fake content can influence public opinion, political debates, and even violate the privacy of individuals, it's important to strike a balance between protecting freedom of expression and preventing potential harm. For actions that may mislead the public, defame individuals, or harm the public interest, there is a need for regulation.

Balancing the protection of freedom of expression with the need to control the potential harms of the abuse of *deepfakes*, it should be strive to find a reasonable balance that

[27] This issue is widely discussed in the literature on the subject, because in the Polish legal system there is no provision directly stating the legal capacity of the deceased author of moral rights. For a broader context, see: E. Szych, *The construction of posthumous substitution of the author in the light of copyright law*, ZNUJ. PPWI, No. 2, 2024, (89-105), pp. 89-105; A.M. Niżankowska, *The Right to the Integrity of the Work*, Warsaw 2007, p. 382.
[28] J. Sieńczyło-Chlabicz, *Prawo do wizerunku a komercjalizacja dóbr osobistych*, 2007, PiP, no. 6, (19-34), pp. 19-34.
[29] Convention for the Protection of Human Rights and Fundamental Freedoms drawn up in Rome on 4 November 1950, subsequently amended by Protocols Nos. 3, 5 and 8 and supplemented by Protocol No. 2. (Journal of Laws. 1993 no. 61 item 284).

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe” (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

simultaneously ensures freedom of expression and protection against the harmful effects of fake content. Otherwise, there is a risk that technology may become a tool for purposes contrary to the foundations of a democratic society. If the infringement of personal rights occurs as a result of speech, the determination whether the infringement is unlawful must be made taking into account the freedom of expression guaranteed in Article 54(1) of the Constitution[30] and Article 10(1) and (2) of the Convention for the Protection of Human Rights and Fundamental Freedoms.

In 2022, Polish jurisprudence has expressed its opinion on this balance[31]. An article in the weekly violated the image, good name and dignity of the plaintiff, presenting him in a humiliating way and suggesting his participation in actions against the authorities - the plaintiff was depicted in a general's uniform with several decorations, reminiscent of a military uniform of martial law. The Court of Appeals partially changed the verdict but upheld that the publication violated the good name and dignity of the plaintiff. A cassation appeal was filed against the judgment of the court of appeal. In the case of a satirical form presenting potentially false data, it should be assessed whether there is a probability that it will mislead the average recipient. This assessment should consider whether it is an exaggeration typical of the satirical genre or whether it can arouse conviction about the truthfulness of the presented content. The key aspect is to understand whether the form of communication reinforces the belief in the authenticity of the information presented. In particular, the court ruled that the satirical nature of a work does not always exclude the unlawfulness of the author's action. The limit of humorous speech is the protection of human dignity, so the only purpose of satire cannot be to insult or humiliate, it should be accompanied by some political or social message. In addition, the court pointed out that a humorous statement in the context of a public debate has a special status and is subject to intensive protection under art. 10 of the European Convention on Human Rights[32] .

[30] Constitution of the Republic of Poland of 1997 (Journal of Laws 1997, No. 78, item 483).
[31] Judgment of the Supreme Court of August 11, 2022, ref. II CSKP 313/22, Legalis no. 2739182.
[32] Ibidem.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe” (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

5. Regulation of deepfakes - risk classification and transparency obligations in EU AI policy.

Recently, a lot of forces have focused on EU regulations on *artificial intelligence*. It is impossible not to mention that *deepfakes* are one of the areas regulated by the AI Act[33]. The need for transparency in the use of artificial intelligence systems, especially in the context of content that may mislead the public, such as deepfakes, is mentioned in recital 134 of the preamble.

The main idea is that entities using such systems must clearly disclose that the content in question has been created or modified by AI to avoid audiences mistakenly believing it to be authentic. However, this requirement does not limit the right to freedom of expression or the right to freedom of the arts and sciences, as guaranteed by the Charter of Fundamental Rights of the European Union. This applies in particular to cases where the content is creative, satirical, artistic, fictional or similar. In such situations, the obligation to disclose that the material has been generated by AI must be fulfilled in a way that does not adversely affect the perception of the work, its usability or quality. In addition, appropriate labelling should also be ensured for AI-generated text content that is published to inform the public about relevant issues, unless the content has been verified by a human or subjected to editorial control and the responsibility for its publication lies with a specific natural or legal person.

Article 3 provides an official definition of: *deep fake* means AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful. Therefore, this definition considers the impression that the material evokes in the recipient to be crucial. What matters is whether the recipient could wrongly consider the content to be true, and not the intention of the creator.

Entities that use artificial intelligence systems to generate or manipulate images, audio or video content in the form of deepfakes are required to disclose that this content has been artificially created or modified. An exception to this obligation applies when such content is lawfully used for the purposes of detecting, preventing, investigating or prosecuting criminal offenders. Where the generated or manipulated content is part of an artistic, creative, satirical, fictitious or similar work, the obligation to disclose the existence of such content is limited to

[33] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

providing adequate information, in a way that does not interfere with the reception or use of the work.

6. **Summary**

Defense strategies are still being developed against deepfakes and educational activities are being carried out to raise awareness of fake news. For example, researchers from IDEAS NCBR have developed experimental ProvenView software, which uses Zero-Knowledge Proofs (ZKP) technology[34] to authorize the authenticity of video, which is to help protect against deepfakes. This technology is to allow verification whether the recording comes from a real person, without the need to disclose the full content of the material. The software is flexible, scalable, and doesn't require intermediaries like electronic signatures. ProvenView is still in development. Currently, video authenticity verification is done after the recording is completed with the prospect of project development for real-time authenticity checking[35] .

The use of *deepfakes* for human rights, individual identity and attitudes to freedom of speech raises a number of important questions. Law and identity are closely linked, especially in the context of deepfakes, where the right to image and privacy is violated. A person's identity includes their integrity and dignity, and deepfakes can cause serious damage to identity through image and audio manipulation. The right to image and privacy is therefore becoming a key element in the protection of identity in the face of technological advances. A key value for the identity of an individual or group is protection against infringement of personal rights, especially in the context of *deepfakes*, where manipulation of multimedia content can lead to serious consequences for the private and social life of the individual. The interaction between identity and human rights in the context of diversity and inclusion becomes more complex in the face of *deepfakes*. This technology can be used to spread disinformation, attack cultural diversity, and cause social conflict. In this context, it is necessary to consider legal aspects that protect against discrimination and promote tolerance. The essence of self-identification and full participation in social and legal interactions becomes even more important in the face of

[34] Zero-knowledge technology is a cryptographic method where one party (the "prover") can prove to another party (the "verifier") that they know a certain piece of information, without revealing the information itself. This ensures privacy and security by allowing verification of data without sharing sensitive details. A common use case is in blockchain or authentication systems, where can be proved an identity or a transaction's validity without exposing personal data. See more at: https://chain.link/education/zero-knowledge-proof-zkp (Accessed: September 20, 2024).

[35] Blog of the Ideas NCBR, 2024. Accessed: 20.09.2024, https://ideas-ncbr.pl/jak-upewnic-sie-ze-rozmawialismy-online-z-prawdziwa-osobanaukowcy-z-ideas-ncbr-opracowali-prototypowe-rozwiazanie/.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe” (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

deepfakes. Individuals must be able to effectively identify themselves and protect their identity from manipulation. At the same time, legal systems must face these challenges and adapt to new forms of threats.

With regard to right to publicity, it should be assumed that after the death of the original rightsholder, his right would be subject to inheritance under general rules and would expire within the statutory period and would become (similarly to copyright or a related right of the performer) hereditary right under general rules. Therefore, in my opinion, one of the solutions would be to assume that the provision of Article 83 of the Civil Procedure Code. Car specs. provides for posthumous protection of the property elements of the right to image. Then, as part of posthumous protection, the heirs of the deceased may file a claim for compensation for commercial use of the image of the deceased. Broadening the legal perspective to include the idea of narrative unity is important because *deepfakes* can influence the construction of social narratives through fake news and images. Implementing narrative unity in legal and social approaches can be critical to maintaining the consistency and integrity of information. Public institutions need to effectively assess identity claims in the context of *deepfakes*. They must act quickly and effectively to protect individuals from manipulation and harm resulting from fake content. At the same time, these institutions should promote transparency and accountability in deepfakes. *Deepfakes* pose many challenges to society and legal systems related to identity protection, human rights, diversity and freedom of expression, so it is worth constantly adapting the legal framework to the dynamic development of this technology in order to effectively protect individuals and society.

Czy RODO i prawo autorskie mogą powstrzymać deepfake? Kompleksowa analiza prawna zjawiska deepfake i odpowiedzialności za deepfake w polskim systemie prawnym

Streszczenie

Artykuł stanowi kompleksową analizę relacji pomiędzy technologią deepfake a kwestiami tożsamości w kontekście polskiego systemu prawnego. W kontekście dynamicznego rozwoju generatywnych systemów sztucznej inteligencji istotnym aspektem jest ochrona tożsamości osób fizycznych. Niniejszy artykuł koncentruje się na dwóch kluczowych obszarach:

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe” (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

technologii wykrywania oraz ramach prawnych, które równoważą wolność słowa z ochroną tożsamości i bezpieczeństwem informacji. Pierwsza część artykułu koncentruje się na roli technologii wykrywania w zwalczaniu deepfake. Przeanalizowano aktualne trendy i postępy w dziedzinie manipulacji treścią. Drugim ważnym aspektem poruszonym w artykule jest analiza ram prawnych regulujących deepfake w Polsce. Omówione zostały kwestie odpowiedzialności karnej i cywilnej związanej z tworzeniem i rozpowszechnianiem deepfake'ów, ze szczególnym uwzględnieniem ochrony wizerunku i związanego z nią prawa autorskiego i praw pokrewnych. Przeanalizowana została również relacja pomiędzy wolnością twórczą a ochroną indywidualnej tożsamości, w tym przez pryzmat polskiego orzecznictwa. Ostatnia część artykułu koncentruje się na aktualnych zmianach w europejskim AI Act (stan na grudzień 2023 r.). Wskazano, w jaki sposób nowe regulacje AI wpływają na kwestie ochrony tożsamości i zwalczania deepfake.

Słowa kluczowe: Deepfake, ochrona tożsamości, sztuczna inteligencja, uczenie maszynowe, ustawa o sztucznej inteligencji, prawa do wizerunku.

Agata Wróbel

A law student at the Faculty of Law and Administration, Cardinal Stefan Wyszynski University in Warsaw, Poland, ORCID: 0000-0003-1809-0334, e-mail: agatawrobel.poczta@gmail.com. Her main research interests are new technology law, intellectual and industrial property law, and civil law and procedure.

Studentka prawa na Wydziale Prawa i Administracji, Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie, Polska, ORCID: 0000-0003-1809-0334, e-mail: agatawrobel.poczta@gmail.com. Głównie zainteresowania badawcze to prawo nowych technologii, prawo własności intelektualnej i przemysłowej oraz prawo i postępowanie cywilne.

References

Legal acts

1. Act of 23 April 1964 Civil Code [ustawa - Kodeks cywilny] (Journal of Laws 2023.1610 unified text of 2023).

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe” (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

2. Act of 4 February 1994 on Copyright and Related Rights [ustawa o prawie autorskim i prawach pokrewnych] (Journal of Laws 2022.2509 unified text of 2022).
3. Act of 6 June 1997 Penal Code [ustawa - Kodeks karny] (Journal of Laws 2022.1138 unified text 2022).
4. Constitution of the Republic of Poland of 2 April 1997 (Journal of Laws 1997, No. 78, item 483)
5. Convention for the Protection of Human Rights and Fundamental Freedoms drawn up in Rome on 4 November 1950, subsequently amended by Protocols Nos. 3, 5 and 8 and supplemented by Protocol No. 2. (Journal of Laws. 1993 no. 61 item 284)
6. Regulation (EE) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5.2016, p. 1–88)
7. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (OJ L, 2024/1689, 12.7.2024)
8. Act of 16 February 2007 on the Protection of Competition and Consumers [ustawa o ochronie konkurencji i konsumentów] (Journal of Laws. 2007 No. 50 item 331 unified text of 2023)

Literature

1. A. Lach, *Kradzież tożsamości,* Prokuratura i Prawo 3, 2012, pp. 29–39.
2. Matlak, *Cywilnoprawna ochrona wizerunku*, Kwartalnik Prawa Prywatnego 13(2), 2004, pp. 317–358.
3. Ziobroń, *Deepfake a prawo karne. Uwagi de lege lata i de lege ferenda dotyczące fałszywej pornografii*, Studenckie Prace Prawnicze, Administratywistyczne i Ekonomiczne 37, 2021, pp. 225–238.
4. A.M. Niżankowska, *The Right to the Integrity of the Work*, Warsaw 2007
5. Blog of the Ideas NCBR, 2024. Accessed: 20.09.2024, https://ideas-ncbr.pl/jak-upewnic-sie-ze-rozmawialismy-online-z-prawdziwa-osobanaukowcy-z-ideas-ncbr-opracowali-prototypowe-rozwiazanie/
6. BuzzFeedVideo, *You Won't Believe What Obama Says In This Video!*, 2018. Accessed 29.12.2023, https://www.youtube.com/watch?v=cQ54GDm1eL0
7. Channel 4 Entertainment, *Deepfake Queen: 2020 Alternative Christmas Message*, 2020. Accessed 29.12.2023, https://www.youtube.com/watch?v=IvY-Abd2FfM
8. D. Shelton, *Capturing The Last of Us: Motion Capture Pipeline*, 2015. Accessed 29.12.2023, https://www.gdcvault.com/play/1021854/Capturing-The-Last-of-Us
9. R.A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, Fordham Law Review 88(3), 2019, pp. 886–938)
10. DiepNep, *This is not Morgan Freeman - A Deepfake Singularity*, 2021. Accessed 29.12.2023, https://www.youtube.com/watch?v=oxXpB9pSETo&t=1s
11. E. Ferenc-Szydełko, *Ustawa o prawie autorskim i prawach pokrewnych. Komentarz*, Warsaw 2021
12. E. Szych, *The construction of posthumous substitution of the author in the light of copyright law*, ZNUJ. PPWI, No. 2, 2024, pp. 89-105.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

13. I. Zgoliński, *Commentary on Article 216*, [in:] V. Konarska-Wrzosek (ed.) *Kodeks karny. Komentarz*, Warsaw 2021, pp.1146–1150.
14. J. Barta, M. Czajkowska-Dąbrowska, Z. Ćwiąkalski, R. Markiewicz, E. Traple, *Prawo autorskie i prawa pokrewne Komentarz*, Wolters Kluwer, 2011.
15. J. Kluza, *Insult in Polish Penal Law. Crime of Art. 216 of the Polish Penal Code in the Perspective of the Functioning of the Judiciary with a Particular Regard to Case Law*, Studenckie Zeszyty Naukowe 20(35), 2017, pp. 60–70.
16. J. Sieńczyło-Chlabicz, *Prawo do wizerunku a komercjalizacja dóbr osobistych*, 2007, PiP, no. 6, pp. 19-34.
17. J. Wilpon, T. Thomson, S. Bangalore, P. Haffner, M. Johnston, Interactions LLC*, Fundamentals of Machine Learning, whitepaper*. Accessed: 20.09.2024, https://www.interactions.com/resources/technology/fundamentals-machine-learning/.
18. K. Rainko, A. Sikora, *Wizerunek w świetle prawa cywilnego a ochrona danych osobowych* [in:] *Ochrona danych osobowych w prawie publicznym*, ed. M. Jędrzejczak, Warsaw 2021, pp. 161-174.
19. K. Szpyt, *Sztuczna inteligencja i nowe technologie (nie zawsze) w służbie ludzkości, czyli cywilnoprawna problematyka rozwoju i popularyzacji technologii deepfake*, [in:] K. Flaga-Gieruszyńska (ed.), *Sztuczna inteligencja, blockchain, cyberbezpieczeństwo oraz dane osobowe. Zagadnienia wybrane*, Warsaw 2019, pp. 75–94.
20. M. Modrzejewska, *Podobieństwo awatarów do znanych osób żyjących* [in:] *Verba volant, scripta manet. Księga jubileuszowa dedykowana Pani Profesor Bogusławie Gneli*, ed. A. Kaźmierczyk, K. Michałowska, M. Szaraniec, Warsaw 2023, pp. 721-731.
21. M. Sewastianowicz, *Deepfake - ofiara realistycznej przeróbki może mieć problem z dochodzeniem swoich praw*, 2023. Accessed 29.12.2023, www.prawo.pl
22. M. Skwarzyński, *Podmiotowe granice wolności słowa w wypowiedziach publicznych*, Law in Action, Civil Cases, vol. 48, 2021, pp. 261–271.
23. M.L. Joshi, *Deepfake: A Systematic Review*, Kalyan Bharati, No. 0976-0822, 2022, UGC-CARE List Group I (73-74)
24. N. Bochyńska, *Klonowanie głosu. Fałszywe próbki mogą oszukać systemy uwierzytelniania*, CyberDefence24.pl, 2023. Accessed: 20.09.2024, https://cyberdefence24.pl/cyberbezpieczenstwo/klonowanie-glosu-falszywe-probki-moga-oszukac-systemy-uwierzytelniania
25. O. Bodanka, *Image violation with the use of deepfake technology – a legal and practical analysis*, Opolskie Studia Administracyjno-Prawne 20(1), 2022, pp. 11–32.
26. PlayStation Australia, *The Last of Us Part I. Motion Capture Trailer*, 2022. Accessed 29.12.2023, https://www.youtube.com/watch?v=4Wd0P5WnO9Q
27. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Accessed: 29.12.2023, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
28. S. Haysom, *People are using face-swapping tech to add Nicolas Cage to random movies and what is 2018*, 2018. Accessed 29.12.2023, https://mashable.com/article/nicolas-cage-face-swapping-deepfakes
29. S. Wieczorek, M. Kubiak, *Ryzyka i szanse wynikające z rozwoju nowych technologii w branży mediowej na przykładzie zjawiska deep fake – analiza prawna*, Monitor Prawniczy 21, 2019, pp. 101–107.
30. V. G. Garcia, *Deep fake – postęp technologiczny a prawo karne*, Acta Iuridica Resoviensia, 44(126), 2024, pp. 39–53.
31. X. Mingchuan Long, M. Zha, *An Overview of Generative Adversarial Networks*, Journal of Computing and Electronic Information Management, 10(3), 2023, pp. 31-36.

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe" (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj

Judgments

1. Judgment of the Court of Appeal in Warsaw of 19 September 2018, ref. VI ACa 528/17, LEX nr 2616080.
2. Judgment of the Supreme Court of August 11, 2022, ref. II CSKP 313/22, Legalis no. 2739182
3. Judgment of the Provincial Administrative Court in Warsaw of 23 Novemeber 2020, ref. II SA/Wa 895/20, Legalis no. 2658998

Licencja Creative Commons „Uznanie autorstwa – bez utworów zależnych 4.0 Międzynarodowe” (CC BY-ND 4.0)
https://czasopisma.uksw.edu.pl/index.php/mj