Deepfake Technology and Current Legal Status of It

Min Liu[1], Xijin Zhang[2*]

[1]*People's Public Security University of China, Beijing, China*
[2]*Key Laboratory of Police Internet of Things Application Ministry of Public Security. People's Republic of China*
liumin1323@163.com, xijinzhang@163.com

Abstract

Deepfake refers to the artificial intelligence technology that splices individual sounds, facial expressions, and body movements into false content with the help of neural network technology. It makes it possible to tamper with or generate highly realistic audio and video contents and make it difficult to identify, which observers fail to distinguish with the naked eye. Therefore, the abuse of deepfake technology will accelerate human beings into the "post-truth era", which will cause a series of social risks, endangering personal legitimate rights and interests, social and public security, and even national security. This paper provides an overview of the main algorithm model—Autoencoder and Generative Adversarial Network—of deepfake, and then points out the existing risks and legal regulation of deepfake technology.

Keywords: *Deepfake, Autoencoder, Generate Adversarial Networks, Legal Regulation*

1 INTRODUCTION

"Deepfake" is a portmanteau of "deep learning" and "fake" [2], which is based on the deep learning algorithm model that can learn independently, especially Generative Adversarial Networks. It first came to the public view in 2017, when a Reddit user named "deepfaker" posted a deepfake video replacing a female star's face with a heroine in a pornographic video. The user's name, "deepfaker", was then used to name the deepfake technology [15]. After that, similar pornographic deepfake videos have gone viral, with many celebrities and even the public becoming victims of pornographic videos. Moreover, some deepfake videos about politicians such as Trump and Obama have also emerged, seriously endangering the national image and diplomatic security.

The rapid development of technology makes the threshold for the use of deepfake lower, which is popularized in a low-cost way and is easily accessible to amateurs so that all individuals may become malicious users or victims of deepfake technology.

This paper begins with an introduction of the technology, which is used to create deepfakes. We then move on to discuss the current harms of deepfake. In the end, we move to explore the current legal and policy status of deepfakes and provide prospects for the regulation of deepfake technology.

2 TECHNOLIGY FOUNDATIONS

Before the emergence of deepfake technology, forgery is usually achieved through the splicing of videos and images. The process of splicing is also a process of covering, by removing, duplicating, shifting, or deleting to achieve the covering and splicing of certain objects [18]. Unlike the splicing of the images and videos, deepfake technology originated from Convolutional Neural Network, which is one of the representative algorithms of deep learning. Initial video image forgery mainly depends on the Auto-encoder Network.

2.1 Autoencoder

Autoencoder is an artificial neural network architecture divided into two parts: encoder and decoder [16]. The encoder encodes and compresses face images by extracting the face features, transforming the image into vector values in the latent space, while the decoder reconstructs the original face according to the face features extracted by the encoder, making the data as close as possible to the input data of the encoder.

It requires two pairs of encoder-decoder to enable source and target image face exchange, and the parameters of two sets of input images are shared between the two encoders, respectively reconstructing the images using different decoders during decoding [22]. The specific operation process is shown in Figure 1.

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 1308-1314, 2023.
https://doi.org/10.2991/978-94-6463-040-4_194

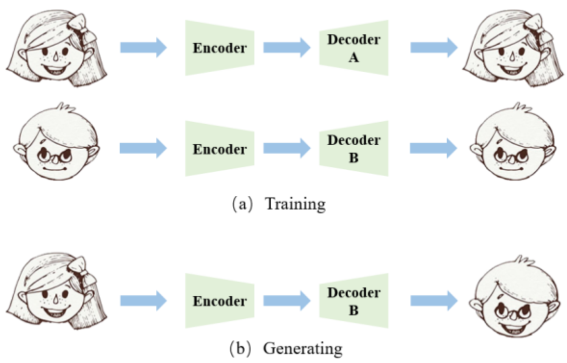

Figure 1. Deepfake generation based on Autoencoder

First, the two encoders extract facial features from the source image and the target image respectively. Then, two different decoders would reconstruct their facial expression. Finally, by exchanging the decoder of the source and target image, where the feature set of the face A is connected to the decoder B, to generated fake image. The newly generated target image has the face features of the source image A while maintaining the face expression and characteristic attributes of the target image B.

However, the Autoencoder network needs to deliberately approximate the probability distribution of the real sample data to improve the fidelity of deepfake, resulting in insufficient network generalization performance and limited generated fidelity.

2.2 Generative Adversarial Network

The GAN technology, as the underlying model of "deepfake", was proposed in Generative Adversarial Networks by Ian J. Goodfellow et al. in October 2014 [13]. Its core idea comes from the two-person zero-sum game in game theory. Traditional deep learning technology is basically a single-level process, but GAN introduces an "adversarial" mechanism, which relies on the repeated creating and detection of internal algorithmic data. GAN is carried out bidirectionally by two sets of deep convolutional neural networks learning in a dynamic, including generator and discriminator.

The "generator", based on the deep learning of the statistical patterns in a data set, generates convincing forged images or videos. The "discriminator" identifies the authenticity of the simulated samples based on the real image, sends the discrimination results back and informs the part to be corrected, the generator then takes a turn, refining the video and eliminating errors. The two are trained in an iterative process, as shown in Figure 2.

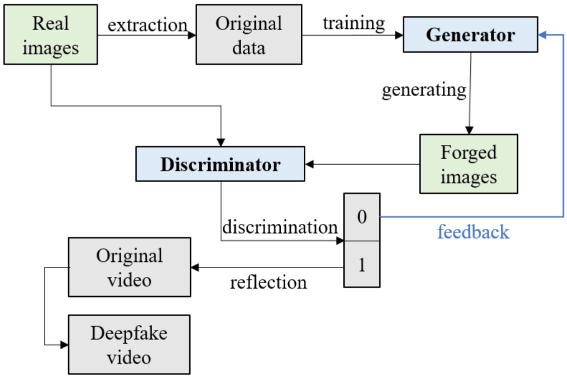

Figure 2. Deepfake generation based on GAN

The objective function formula of the GAN training is given by equation 1:

$$\min_{G}\max_{D}\ (D,G) = E_{x\sim P(x)}[\log D(x)] + E_{z\sim P(z)}[\log(1-D(G(z)))] \quad (1)$$

Among them, G is the generator, D is the discriminator and E represents expectation of the distribution function. The generator maps the input data z satisfying the random distribution from the input space, recorded as $x=G(z)$, and then maps x to $D(x)$ through the discriminator, and solves the objective function by taking the expectation.

The generator and discriminator are trained in a min-max way method [13]. The minimum value 0 represents the fake output and the maximum value 1 represents the authentic output. The cost function of G wants $V(D, G)$ to be as small as possible, while D wants $V(D, G)$ to be as large as possible to form a game between the two [21]. D tries to get close to 1 to create an authentic deepfake output. If we get $D(G(z))$ ~0, then this means that G cannot fool D. After repeated training and detecting, until the discriminator cannot accurately identify the difference between the generator's outputs and the original data set, which means that the generated data and the real data have the same distribution, the whole forgery process comes to an end, leading to an incredibly realistic video that can deceive the eyes of most people.

3 HARMS OF DEEPFAKES

As the deepfake technology becoming increasingly accessible to non-professionals, leading to a surge in the number of fake audio and video products. According to the 2018 report "Malicious Use of Artificial Intelligence: Prediction, Prevention and Mitigation" released by the Institute for Future Life [3], there is a risk of malicious use of AI, and deepfake technology is one of them.

3.1 The Impact on Individuals

Data from The State of Deepfakes, Landscape, Threat and Impact [1] shows that ninety-six percent of deepfake videos were pornographic videos, and the victims were absolutely women. The first use of GANs was to create deepfake pornographic videos, especially revenge porn and celebrity deepfakes. These pornographic deepfakes cause substantial injuries to women, not only workplace discrimination, emotional and reputational harm, but also the sexual exploitation or even the death and rape threats.

3.2 The Impact on Society

Deepfake technology will also further blur the boundary between truth and illusion, causing a crisis of trust in the whole society. In March 2021, the Hongkou District People's Procuratorate of Shanghai Municipality in China prosecuted a large false Value Added Tax Invoice. The criminal suspect forged action videos including nodding, shaking head, blinking, and opening mouth through the technical processing of other people's high-definition profile pictures and ID card information, to crack the face recognition technology and falsely issue ordinary VAT invoices.

Deepfake technology could be used to spread misinformation division and create social unrest. In 2018, for instance, more than twenty people across India were violently killed because of the rumors of kidnapping young children or involving other crimes on WhatsApp [6].

Deepfake technology also poses threats to judicial practice and the legal system. Artificial intelligence technology is more and more used in the courts, if the detection technology cannot keep up with the pace of deepfake technology, it may cause misjudgment of cases, seriously affecting the judicial justice and the interests of the victims.

3.3 The Impact on Nations

There have been many concerns about the political deepfake videos to interfere with elections. Similar videos were used to target President Joe Biden in 2020 election in the United States [9]. Deepfakes would also disrupt diplomatic relationships. The diplomatic crisis in the Middle East is considered to be related to false information events [14].

Many fake videos of politicians and national leaders have been circulating on social media. Although the videos are now more just laughed off as entertainment, as the deepening of deepfake technology, these fake videos will become more and more realistic and more difficult for ordinary people to distinguish them, so the destruction of political figures will become more serious.

4 EXISTING LEGISLATION OF DEEPFAKE

For the risk of technology, the usual logic is "Defeat magic with magic"—with technology. However, the rate of technological progress is often faster than the speed at which the technology can be broken. So, it is urgent and necessary to regulate deepfake technology by means other than technology. On March 7, 2020, a symposium—"When Seeing Isn't Believing: Deepfakes and the Law"—was held in New York, focusing on the legal and regulatory response to deepfakes [20].

4.1 The United States

The United States was the first country to respond to artificial intelligence technology. In December 2018, the U.S. Congress passed Malicious Deep Fake Prohibition Act of 2018 [17], which was the first act to define the Deepfake. DEEPFAKES Accountability Act was introduced in June 2019 [12]. However, it has been challenged and opposed by the public for its vague definitions and a potential conflict with the First Amendment to the United States Constitution [11]. In the same year, the Congress proposed the Deepfake Report Act of 2019 [5], requiring the U.S. Department of Homeland Security to regularly issue the evaluation reports on deepfake technology.

In addition, some states respond quickly to the improper use of "deepfake", especially on "pornographic videos" and "political elections".

Table 1. Legislation of the United States

	Regulations	Time	Content
Federal legislation	Malicious Deep Fake Prohibition Act	In December 2018	set up reporting systems
	DEEPFAKES Accountability Act of 2019	In June 2019	label the altered media
	The Deepfake Report Act of 2019	In June 2019	issue reports on deepfake technology
Virginia	Unlawful Dissemination or Sale of Images of Another Person	In July 2019	nonconsensual deepfake pornography

Texas	Tex. SB 751	In September 2019	election
California	Calif. AB-602	In February 2019	nonconsensual deepfake pornography
	Calif. AB-730	In October 2019	election
	Calif. AB-1280	In September 2021	election and nonconsensual deepfake pornography
Washington	SB 6280 Act	In March 2020	face recognition
New York	N.Y. A08155, S0587-B	In November 2020	nonconsensual pornography, "digital replica" and commercial exploitation
Massachusetts	An Act to Protect Against Deep Fakes Used to Facilitate Criminal or Torturous Conduct	In January 2019	establish liability on "facilitate criminal or tortious conduct"

4.2 The European Union

The EU has not issued special legislation on "deepfake" but has adopted a series of regulations and programs to incorporate deepfake into the regulatory framework, limiting the application of deepfake from disinformation governance, individual information protection and artificial intelligence regulation.

In April 2018, the European Commission published a long open letter entitled Tackling online disinformation: a European Approach, putting forward some principles to avoid information publishers illegally manipulate public opinion [8]. In May 2018, the European Union formally implemented the General Data Protection Regulations. This regulation set strict rules on the use of deep synthesis technology, protecting personal data such as images of citizens that may be used for deepfake [19]. In June 2018, the European Council adopted the EU Code of Practice on Disinformation, actively promoting self-regulation of the industry and consciously restricting and controlling the illegal content of "deepfake"[7].

Table 2. Legislation of the European Union

Regulations	Time	Content
General Data Protection Regulations	In May 2018	personal data
Tackling online disinformation: a European Approach	In April 2018	illegally manipulate public opinion
Code of Practice on Disinformation	In June 2018	advocate self-regulation of platforms
Ethics Guidelines for Trustworthy Artificial Intelligence	In April 2019	privacy and data management

4.3 China

China also does not carry out special legislation on deepfake, but standardizes and restricts the creation, release and dissemination of deepfake information from the perspective of protecting citizens' right of portrait, reputation, and safeguarding national security and social security. Moreover, its legal regulations focus on the obligation of labelling.

Unfortunately, there are no punitive provisions for violations of the labelling obligation, which makes the declaration of the provisions more meaningful than the practical value, resulting in the absence of legal protection.

Table 3. Legislation of China

Regulations	Time	Content
Data Security Management Measures (draft)	In May 2019	the obligation of labelling
The Regulations on the Administration of Online Audio and Video Information Services	In January 2020	
Network Information Content Ecological Governance Regulation	In March 2020	
The Civil Code of the People's Republic of China	In January 2021	personal right

5 CONCLUSION

Today, video has been a relatively reliable source of information, but once "deepfake" becomes more popular, the value of any video—whether true or false—inevitably falls, because there is no reliable way to determine whether a video is forged or not.

The law is only a kind of passive post-event relief. Although it can exert certain constraints on the dissemination of false information in certain fields and specific scenes, it is ineffective to the negative impact already caused, and the social credibility of social media often gradually weakens with the development of emergencies.

Therefore, prior prevention and in-process control are particularly crucial. The most critical links are the creators, platforms and audiences, what should be done is to delimit the application boundaries of new technology by ethical norms, guide the development direction of new technology with industry self-discipline and strengthen the education of critical thinking of the public.

More importantly, such regulatory responses will not be so efficient without significant technical expertise, so we need both lawyers and technologists to tackle this problem [4]. As the deepfake technology becomes more and more mature, the corresponding detection will be more and more advanced. It will be a never-ending race, which Doermann compared to a "cat-and-mouse game" [10].

ACKNOWLEGMENTS

The work described in this paper was supported by Key Laboratory of Police Internet of Things Application Ministry of Public Security. People's Republic of China.

REFERENCES

[1] Ajder H. Deepfake Threat Intelligence: A statistics snapshot from June 2020 [J]. 2020.

[2] Brandon J. Terrifying high-tech porn: creepy 'deepfake' videos are on the rise[J]. Fox news, 2018, 20.

[3] Brundage M, Avin S, Clark J, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation [J]. arXiv preprint arXiv:1802.07228, 2018.

[4] Chesney R, Citron D K. 21st century-style truth decay: Deep fakes and the challenge for privacy, free expression, and national security [J]. Md. L. Rev., 2018, 78: 882.

[5] Deepfakes Report Act of 2019, https://www.congress.gov/bill/116th-congress/house-bill/3600/.

[6] Donie O'Sullivan. House Intel chair sounds alarm in Congress' first hearing on deepfake videos [EB/OL]. https://edition.cnn.com/2019/06/13/tech/deepfake-congress-hearing/index. html.

[7] European Commission. EU code of practice on disinformation [EB/OL]. https://www. hadopi.fr/sites/default/files/sites/default/files/ckeditor_files/1CodeofPracticeonDisinformation.pdf.

[8] European Commission. Tackling Online Disinformation: A European Approach[J]. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. COM/2018/236, 2018: final.

[9] Generally 'FBI Chief Calls Capitol Attack Domestic Terrorism and Rejects Trump's Fraud Claims', The Guardianhttps://www.theguardian.com/us-news/2021/jun/10/capitol-attackfbi-christopher-wray-congress.

[10] Hao K. Deepfakes have got congress panicking. this is what it needs to do[J]. online] MIT Technology Review, 2019, 20.

[11] Hayley Tsukayama, India Mckinney, Jamie Williams. Congress Should Not Rush to Regulate Deepfakes [EB/OL]. https://www.eff.org/deeplinks/2019/06/congress-shouldnot-rush-regulate-deepfakes.

[12] H.R.3230-Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, http://www.congress.gov /bill /116th-congress/house-bill /3230.

[13] Ian Goodfellow, et al. Generative adversarial nets. Advances in neural information processing systems, 2014, 2672-2680.

[14] Krishnadev Calamur, Did Russian Hackers Target Qatar?, THE ATLANTIC [EB/OL]. https: //www.theatlantic.com/news/archive/2017/06 /Qatar-russian-hacker-fake-news/529359 /.

[15] Leo Kelion. Deepfake Porn Videos Deleted from Internet by Gfycat [EB/OL]. http://www. bbc.com/news/technology-42905185.

[16] Ramadhani K N, Munir R. A Comparative Study of Deepfake Video Detection Method [C]//2020 3rd International Conference on Information and Communications Technology (ICOIACT). IEEE, 2020: 394-399.

[17] S. 3805, 115th Cong. (2018).

[18] Thakur R, Rohilla R. Copy-move forgery detection using residuals and convolutional neural network framework: a novel approach[C]//2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC). IEEE, 2019: 561-564.

[19] Voigt P, Von dem Bussche A. The EU General Data Protection Regulation (GDPR)[J]. A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017, 10(3152676): 10.5555.

[20] Yamaoka-Enkerlin A. Disrupting disinformation: Deepfakes and the Law[J]. NYUJ Legis. & Pub. Pol'y, 2020, 22: 725.

[21] Yuxuan Bao, Tianliang Lu, Yanhui Du, Overview of Deepfake Video Detection Technology [J]. Computer Science, 2020,47(09):283-292.

[22] Yuzhi Zhang, Ruifang Wang, Liang Zhu, et al. The Review of Generation and Detection Techniques for Deepfakes [J]. Journal of Information Security Research, 2022,8(03):258-269.