Artificial Intelligence Governance - Taking Deep Fake as an Example

1.Introduction

With the increasing maturity of the use of neural networks, the application of artificial intelligence technologies is becoming more and more widely used. Among them, through the automated editor and convolutional neural network technology, the threshold of the technology of copying films is not very high. In November 2017, some films that superimpose the faces of social celebrities on pornographic film actors/actresses appeared in the American social networking platform, Reddit. These types of films analyze the faces of specific socialites through deep learning algorithms and superimpose their faces on the films, making them look as if the films were taken by the socialites themselves. This technology was released by developers in 2018 and was made into an app for public use. At present, such technology is generally referred to as "deep fake" internationally, and it is believed that it may contribute to the speedy invention and distribution of false information existing throughout the Internet nowadays, which has attracted the attention of legislators worldwide. As it uses fake images or films automatically generated by Deep-learning technology, it involves both dimensions of fake information prevention and artificial intelligence governance. The purpose of this paper is to observe the relevant policies, legal measures and related guidelines or principles of the international community in response to issues of deep fake and artificial intelligence governance, and to examine whether the current legal system in Taiwan can cope with the impact of deep fake so as to provide feasible recommendations.

2.Ethics Rules for Artificial Intelligence

In the governance of artificial intelligence, the European Union introduced the “Ethics Guidelines for Trustworthy AI” on April 8, 2019 to establish a framework for supervising artificial intelligence in order to make artificial intelligence trustable.

The guidelines first points out that Trustworthy AI requires three key characteristics: (1) it should be lawful: complying with all applicable laws and regulations; (2) it should be ethical: ensuring adherence to ethical principles and values; and (3) it should be robust: both from a technical and social perspective, to avoid AI from inadvertently causing harm.

Fundamental Rights are the basis of trustworthy AI. In order to comply with the above-mentioned basic human rights and to make AI reliable, their expert group believes that AI should abide by four ethical principles, including: (1) respect for human autonomy; (2) prevention of harm; (3) fairness; and (4) explicability. The four ethical principles are also transformed into the seven specific measures: “human agency and oversight”, “technical robustness and safety”, “privacy and data governance”, “transparency”, “diversity, non-discrimination and fairness”, “societal and environmental wellbeing impact evaluation” and “AI accountability”. To facilitate the true implementation of self-assessment for application developers, the Guidelines devise the Trustworthy AI Assessment List in Chapter 4 for the reference of the enterprise.

3.Counter measures Against the International false messages

In response to the prevention of false messages, the two parties in the United States also jointly proposed in 2018 the Malicious Deep Fake Prohibition Act of 2018 to amend the relevant provisions of fraud in the criminal law. This bill amends Chapter 47 of the United States Code by adding Section 1041 with regard to fraud in connection with audiovisual records. It treats the use of deep fake as a criminal offence and defines deep fake as “audiovisual record created or altered in a manner that the record would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual”. It shall be unlawful to, using any means or facility of interstate or foreign commerce, to create, with the intent to distribute, a deep fake with the intent that the distribution of the deep fake would facilitate criminal or tortious conduct; or distribute an audiovisual record with actual knowledge that the audiovisual record is a deep fake, and the intent that the distribution of the audiovisual record would facilitate criminal or tortious conduct. Any person who violates the above may be sentenced to imprisonment for more than 2 years but less than 10 years. However, the bill is currently put on hold without being further reviewed.

In addition, in order to properly cope with the danger of deep fake, on June 28, 2019, the two parties in the US Congress jointly proposed the bill - "To require the Secretary of Homeland Security to publish an annual report on the use of deep fake technology, and for other purposes”, which may be cited as the "Deepfakes Report Act of 2019". This bill requires the Department of Homeland Security to conduct research on deep fake and related issues, produce an annual report, and to request it to assess the direction of addition or revision of relevant laws and regulations. Moreover, the US senators from both parties also proposed on June 12, 2019 the bill- “Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019”, which may be cited as “DEEP FAKES Accountability Act”. This Act is the same as the Act of 2018, both of which treat the use of deep fake as a fraudulent act by adding section 1041 to Chapter 47 of the United States Code. However, this Act does not directly define deep fake, but rather define such a type of technology as “advanced technological false personation record”, and require such records to comply with:

(1) DIGITAL WATERMARK: Any advanced technological false personation record which contains a moving visual element shall contain an embedded digital watermark clearly identifying such record as containing altered audio or visual elements.

(2) AUDIOVISUAL DISCLOSURE shall comply with the following principles:

A. clearly articulated verbal statement that identifies the record as containing altered audio and visual elements, and a concise description of the extent of such alteration; and

B. an unobscured written statement in clearly readable text appearing at the bottom of the image throughout the duration of the visual element that identifies the record as containing altered audio and visual elements, and a concise description of the extent of such

alteration.

(3) VISUAL DISCLOSURE shall comply with the following principles: Any advanced technological false personation records exclusively containing a visual element shall include an unobscured written statement in clearly readable text appearing at the bottom of the image throughout the duration of the visual element that identifies the record as containing altered visual elements, and a concise description of the extent of such alteration.

(4) AUDIO DISCLOSURE shall comply with the following principles: Any advanced technological false personation records exclusively containing an audio element shall include, at the beginning of such record, a clearly articulated verbal statement that identifies the record as containing altered audio elements and a concise description of the extent of such alteration, and in the event such record exceeds two minutes in length, not less than 1 additional clearly articulated verbal statement and additional concise description at some interval during each two-minute period thereafter.

According to the bill, those who violate the above requirements shall be subject to legal responsibilities. In criminal liabilities, whoever knowingly violates the above requirements and (1) with the intent to humiliate or otherwise harass the person falsely exhibited, provided the advanced technological false personation record contains sexual content of a visual nature and appears to feature such person engaging in such sexual acts or in a state of nudity; (2) with the intent to cause violence or physical harm, incite armed or diplomatic conflict, or interfere in an official proceeding, including an election, provided the advanced technological false personation record did in fact pose a credible threat of instigating or advancing such; (3) in the course of criminal conduct related to fraud, including securities fraud and wire fraud, false personation, or identity theft; or (4) by a foreign power, or an agent thereof, with the intent of influencing a domestic public policy debate, interfering in a Federal, State, local, or territorial election, or engaging in other acts which such power may not lawfully undertake, may be sentenced to imprisonment for not more than 5 years. In civil liabilities, any person who violates the above requirements may be subject to a civil penalty of up to US$150,000 per record or alteration, as well as the compensation for the damage, if any.

In addition to the United States, the United Kingdom also launched the "Online Harms White Paper" in April 2019, which will establish a new "Online Safety" control structure to respond to false messages and underage pornographic videos, deep fake and online drug trafficking and so on.

The report points out that the new network security control framework will clarify the legal obligations of the Internet company to make the company assume more security responsibilities and avoid the harm caused by the content or actions generated by the service provided, and establish an independent regulatory agency supervising and implementing the relevant legal policies. The regulatory authority should provide relevant guidelines for compliance with the new obligations. If the company is unwilling to comply with the relevant guidelines, it must bear the burden of proof and prove that its alternative measures can achieve more effectively for the purpose of protecting the Internet users. In addition, the framework will also include elements of “Transparency, Trust, and Accountability”. The competent authority will be given the right to request an annual transparency report be submitted by the company, which the report should indicate the relevant harmful contents appeared on its platform, explain how it is handling with the problem, and publish the report on the website. Furthermore, the competent authority will have the right to request additional information from the Internet company, such as how its algorithm works.

In response to false messages, the report points out that current Internet companies have begun to conduct research on the prevention and control methods of fake news dissemination, including: (1) through the terms of service, users are not allowed to distort their identity on social software to spread false messages. (2) developing relevant tools to detect suspicious, false or junk accounts; (3) using automated artificial intelligence to delete or remove fake accounts; and (4) collaborating with independent fact verifying platforms. However, in the future, the government hopes that the guidelines and related policies proposed by the competent authorities must further include the following matters: (1) The company shall clarify its definition of false information in its terms of service, and state its expectations of users, and the possible penalties to users who violate the company policy; (2) The company should adopt the relevant countermeasures to deal with users with distorted identities who disseminate false messages; (3) The visibility of the disputed content currently under the fact-verifying inspection shall be reduced; (4) The fact-verifying service shall be used, especially during the election period, for fulfilling the obligation of fact verification; (5) Promote authoritative news sources; (6) Promote news circulation from different perspectives, rather than only reinforce the messages of people's existing views; (7) Users should be able to recognize that they are interacting with automated accounts and should ensure that the dissemination of automated accounts information is not abused; (8) Promote the transparency of political advertising to comply with the norms of the UK electoral law; (9) Companies should ensure that users may mark the content that they believe to be false news by themselves and let them know that the company is targeting false news for countermeasures to be taken; (10) The procedures for publishing information should be open and transparent so that the public can assess the effectiveness of the company’s response to false information, and further support the relevant research on online false message activities; (11) The relevant procedures and measures should be taken to continuously monitor and evaluate the effectiveness of the processing flow of fake messages.

From the above-mentioned relevant international legal policy observations, it can be found that international measures related to deep fake can be classified into the following items:

(1) Establish an independent fact-verifying unit.

(2) Improve the transparency of information sources.

(3) Improve the oversight responsibility of the online platform for the messages appeared on such a platform.

(4) Deep fake is to be treated as an independent criminal act and its criminal, civil and administrative responsibilities are to be clearly regulated.

(5) On the technical level, relevant artificial intelligence tools are being developed to respond to this issue. For example, the American startup company, Deeptrace, has begun to conduct research and develop deep fake identification technology to identify the authenticity of the films.

Yu, He-Chien
Legal Researcher

Release : 2019/10

Tag