Research on "Deepfake" Threats and Security Strategies

Dingxiang Defense cloud business Security Intelligence Center

2024.03

I. The "Deepfake" technology

The term "Deepfake" is a technology that uses artificial intelligence technology to replace a human face on another person's face.

The "Deepfake" technology involves a variety of technologies and algorithms that work with each other to generate very realistic images or videos. Combining the false content of "Deepfake" with the elements of real information, it can be used to forge identity, spread false information, make false digital content, and conduct all kinds of fraud.

1.1 the composition of the "Deepfake"

1.1.1 Face recognition and key point detection

Face recognition and key point detection are used to identify and locate faces, which is the basis of "Deepfake" technology. It is mainly used to identify faces in images or videos, and to locate key points of faces, such as eyes, nose, mouth, etc. This information can be used to extract faces from images or videos and synthesize them with other images or videos.

among, Face recognition algorithm includes Eigenfaces, Fisherfaces, Local Binary Patterns (LBP), Subspace-based face recognition for Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), And DeepFace, FaceNet, Dlib, etc., based on deep learning for face recognition; The key point detection algorithm includes Active Shape Model (ASM), Constrained Local Model (CLM), Cascaded Shape Regression (CSR), DeepPose, etc.

1.1.2 Image / video synthesis techniques

Image / video synthesis technology combines different faces, expressions and movements into target images or videos, which is the core of "Deepfake" technology. It is mainly used to synthesize different faces, expressions and movements into target images or videos.

Common image / video synthesis technologies include Poisson Blending and Seamless Cloning for pixel-based image / video synthesis, and FaceSwap and DeepFaceLab for feature-based image / video synthesis.

1.1.3 generative adversarial network (GAN)

Generative adversarial network (GAN) is a deep learning technology that can generate very realistic images or videos. GAN consists of two neural networks: generator and discriminator. The

goal of the generator is to generate highly realistic fake images, videos and sounds. It is an important tool of "Deepfake" technology, which includes DCGAN, ProGAN and StyleGAN.

1.1.4 Enhance the effect technology

Techniques that can enhance the effect of "Deepfake", including for synthesizing fake WaveNet, Tacotron, and for creating realistic 3D models like Shape from Shading and ructure from Motion.

1.2 Hazards of the "Deepfake"

In January 2024, Dingxiang Defense Cloud Business Security Intelligence Center issued a warning that AI is becoming a new threat, and that attackers using AI technology will bring unprecedented risks to enterprises and individual users, and listed five trends of business fraud risk in 2024.

Figure 1-1: "AI face change" is technically called "Deepfake"

1.2.1 "Face change" video and image

In January 2024, an employee of a Hong Kong multinational company was defrauded of HK $200 million. The employee was invited to a video conference after receiving an email disguised as a company headquarters CFO. During the meeting, except for the employee, everyone else made fake images for the "Deepfake" technology. The fraudsters instructed the employee to transfer HK $200 million to five bank accounts within a week.

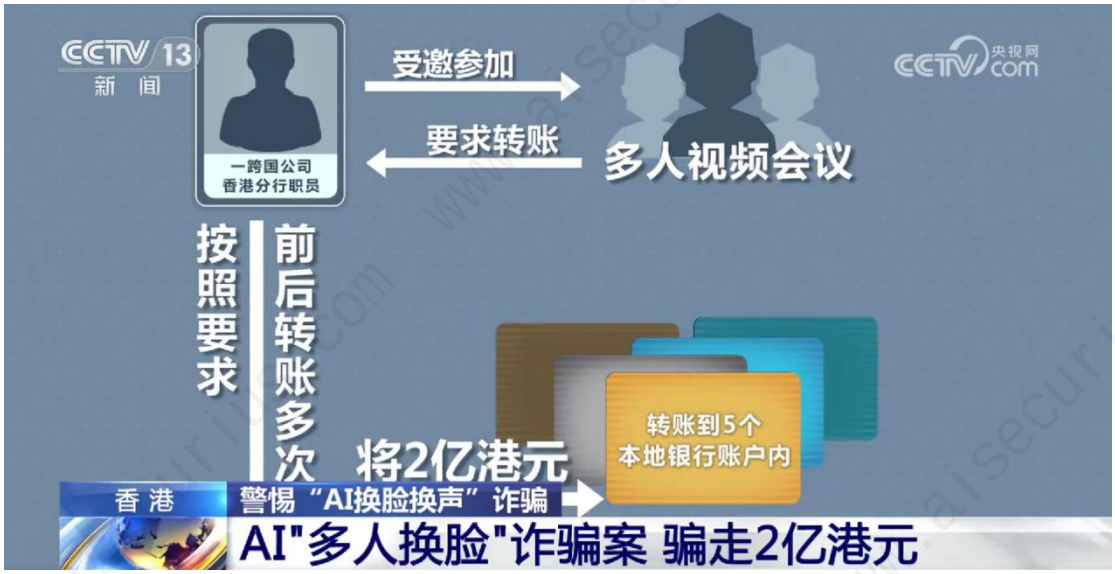

Figure 1-2: Hong Kong HK $200 million was cheated in the "AI face change" case in Hong Kong

In December 2023, an international student was "kidnapped" abroad, his parents were "kidnappers" for a ransom of 5 million yuan, and he received a video of a "meat ticket" being controlled and harmed. Through on-site investigation and international police cooperation, five hours later, the involved student Xiao Jia was successfully rescued —— not from the kidnappers, but at the entry and exit port of the university country. The truth is also revealed: the "kidnapping case" that let Xiao Jia and his parents experience the shocking moment, is actually a scam carefully laid by the fraudsters, which is a typical "virtual kidnapping" fraud for the overseas students and their families.

Fraudsters fake videos and images, replace photos of strangers with the faces of colleagues, and then use remote video calls to attract and persuade the victims to share confidential information (such as credentials) and manipulate the victims into unauthorized financial transactions.

A KPMG report showed that "Deepfake" video available online grew 900% year on year. According to a bandeepfakes report, "Deepfake" pornography accounts for 98 percent of all online "Deepfake" videos, almost all for women.

1.2.2 Synthetic human audio

In November 2023, police in Baotou city, Inner Mongolia Province, reported a fraud case. The fraudsters through the "Deepfake" technology disguised as a friend of Mr.Guo, the legal representative of a technology company in Fuzhou, and then contacted Mr.Guo through the WeChat video. After gaining Mr.Guo's trust, the fraudsters pretended to need to borrow Mr.Guo's account, successfully defrauded Mr.Guo 4.3 million yuan.

The cheater observed and collected the victim for a long time, and had a clear

understanding of the victim's family, work, social relations, whereabouts, track, habits, work and rest, and then used forged videos and voices to carry out directional fraud.

Figure 1-3: The "Deepfake" technology can clone anyone's voice

It's very easy to clone anyone's voice with "Deepfake". Fraud through public channels (for example, video clips on social media, news media reports, public activities of speech conversation, etc.) to extract voice samples, only 30 seconds to 1 minutes of samples, can produce highly realistic sound cloning, and then can be used to leave voice email or let people participate in real-time dialogue, further blurred the boundary between reality and deception.

The fake audio of "Deepfake" poses an growing threat to voice-based authentication systems. Individuals with widely available voice samples, such as celebrities and politicians, are particularly vulnerable to this strategy. According to the survey, 37% of the organizations experienced the voice fraud of "Deepfake" in 2022.

1.2.3 Fake text information

Based on "Deepfake", making fake news, fake news, fake accounts, etc., fraud, extortion, framed, defamation and other illegal acts and cases are common. It is not only harmful to social and economic security, but also related to the safety of corporate personal reputation and personal property.

Figure 1-4: All kinds of "Deepfake" false information is rampant

For example, with businesses already losing billions of dollars a year on email phishing fraud, "deep forgery" makes email phishing attacks more dangerous because fraudsters can use their identities to look more credible. It is easier to create fake corporate executive profiles and use them to lure employees. Phishing emails also pose as employees of key customer companies and often use phrases like "request" "payment" and "urgent" to entice recipients to click.

The "Deepfake" of text information refers to written content that appears to be created by a human person. In particular, the popularity of social media is a major factor in the widespread use of deep textual forgery, in which posts can be used as part of the response of social media manipulation activities or robot generation. Their purpose is often to spread fake news and disinformation on a large scale, creating a deceptive perception that many people on various media platforms share the same beliefs.

2. Common "Deepfake" fraud type

2.1 Fake accounts to spread rumors

With "Deepfake" technology, fraudsters are able to create and distribute highly customized and convincing content that can target individuals based on their online behavior, preferences, and social networks and can be seamlessly integrated into user feeds, facilitating rapid and widespread dissemination. This makes cybercrime more efficient and challenging for users and platforms.

Identifying fake accounts for "Deepfake" is a challenging task because they often involve combinations of real elements (such as real addresses) and fabricated information. This makes detection and prevention extremely difficult. The detection effort is further complicated by using legitimate components and false details. Moreover, because these fraud sexual identities lack a previous credit record or associated suspicious activity, it is difficult to identify them through a conventional fraud detection system.

Figure 2-1: The stranger behind the account number

2.2 posing as acquaintances to commit fraud

In January 2024, employees of Hong Kong multinational companies were cheated of HK $200 million, and the fraud cases reported by police in Baotou, Inner Mongolia in November 2023 were both fake acquaintances.

The "Deepfake" technology allows fraudsters to easily imitate the video and voice of the target person. These fake video and audio can imitate not only sound, but also intonation, accent and speaking style. In order to further increase the credibility of the fraud, the fraudsters will also obtain the sensitive information of the victims (such as work style, living habits, travel trends, etc., etc.) on social media and public channels, to prove the authenticity of their imposter identity, making it difficult for the victims to distinguish the truth from the false. After obtaining the trust of the victim, and then to the victim of financial funds, trade secrets and other fraud.

2.3 Fake your identity to apply for bank loans

According to the McKinsey Institute, synthetic identity fraud has become the fastest growing type of financial crime in the United States and is on the rise globally. Indeed, synthetic identity fraud, accounts for 85% of all current fraud behaviors. Moreover, the UK GDG study shows that more than 8.6 million people in the UK use false or other identities to obtain goods, services or credit.

For financial institutions, the "Deepfake" fraud is even more worrying.scammers use "Deepfake" technology to fake information, sounds, videos, pictures, and then combine real and false identity information to create new false identities to open bank accounts or make fraud purchases. What's more, fraudsters can use the "Deepfake" technology to learn different banking businesses and processes, and then quickly launch a fraud on different banks at the same time.

In the future, the bank for the user's credit application, perhaps not only to evaluate " he / she is suitable for credit 100,000? Or $200,000?", also need to tell" this loan applicant is human? Or Artificial intelligence?", 92% of banks are expected to face a" Deepfake " fraud threat.

2.4 Make "phishing" even more difficult to identify

Phishing has long been a prominent topic in security, and despite such forms of fraud for decades, it is still one of the most effective ways for fraud to attack or penetrate organizations. Fraudsters based on social engineering principle, by email, website, and phone calls, SMS and social media, using human nature (such as impulse, discontent, curiosity), as trusted entities, induce the victims click on false links, download malicious software, induced transfer funds, provide sensitive data such as account password behavior.

With the development of technology, Internet telephers are also changing their strategies, especially with the help of AI, where fraudsters use "Deepfake" technology to deceive victims, making phishing attacks more complex and extremely difficult to detect. In 2023, the deep forgery phishing fraud incidents surged by 3,000%.

2.5 pretend to be others remote interview entry

In July 2022, the FEDERAL Bureau of Investigation (FBI) warned that a growing number of fraudsters are using "Deepfake" technology to pose as job seekers in remote job interviews to defraud salaries and steal business secrets.

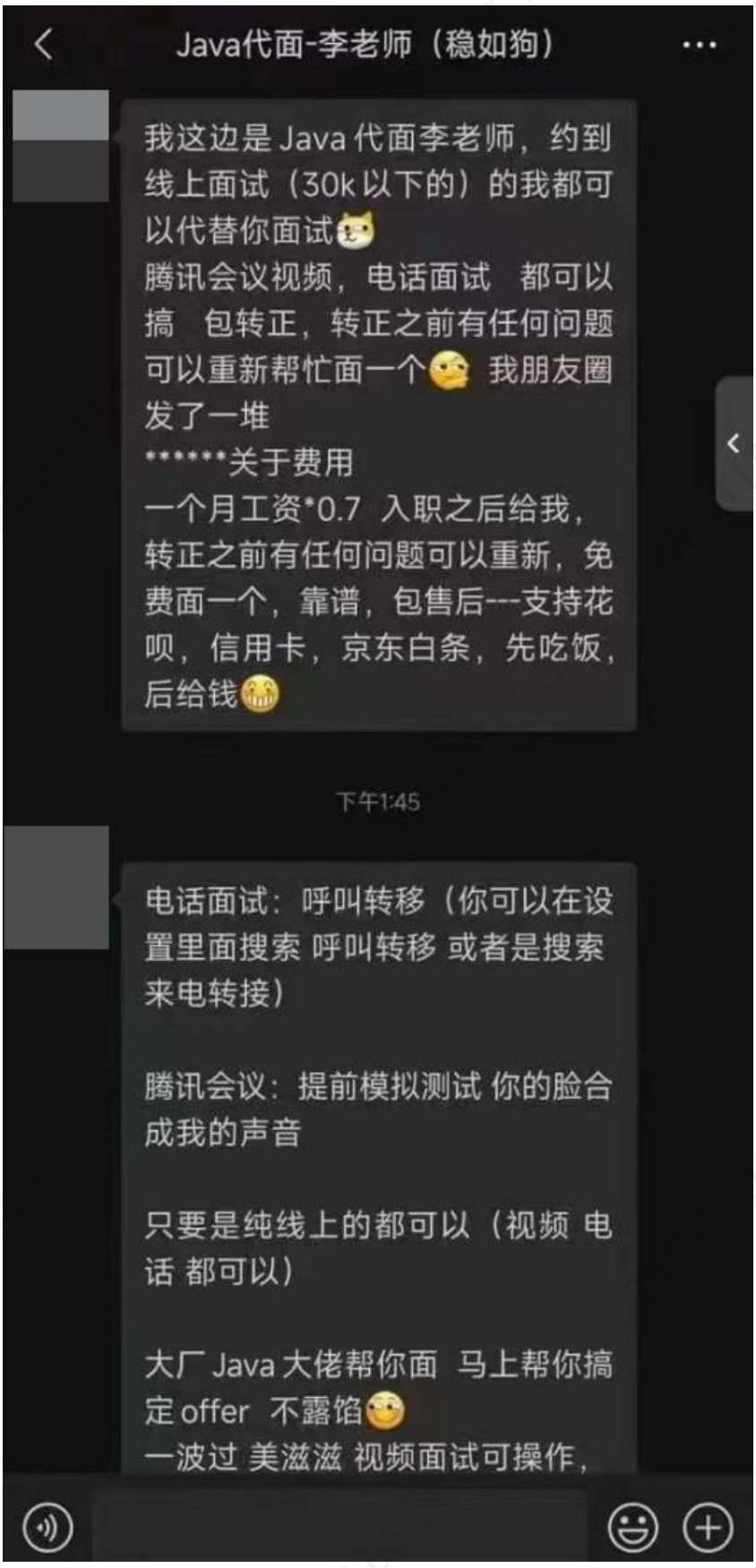

Figure 2-2: Remote interview by posing as someone else

The FBI did not specify the ultimate goal. But the agency noted that the fake interviewers, if successful and hired, will be allowed to successfully access sensitive data such as "customer PII (personal identity information), financial data, corporate IT database and / or proprietary information."

The FBI also said stolen personal information from some of the victims of the scam had been used to conduct remote interviews and had been used to conduct pre-employment background checks with other applicants' profiles.

2.6 fake login to steal the bank balance

On February 15,2024, the foreign security company Group-IB announced that it had found a malware named "GoldPickaxe". The iOS version of the malware lured users into face recognition, submitting identity documents, and then making deep forgery based on the user's face information.

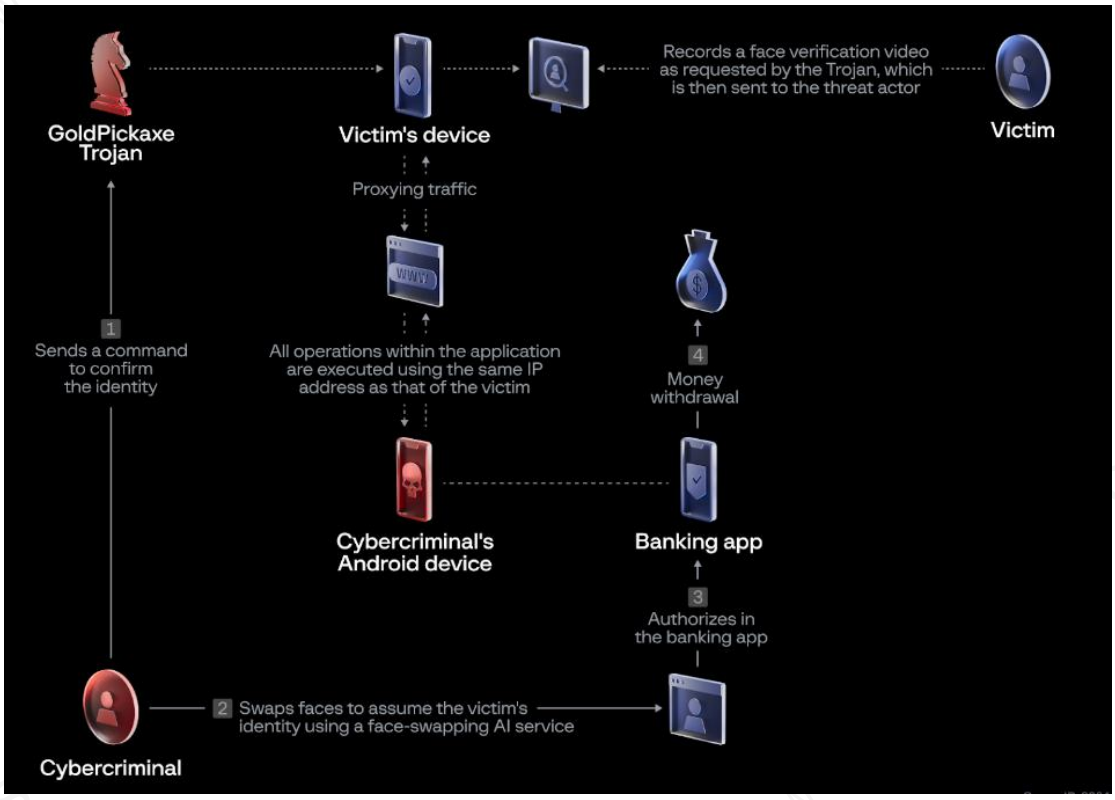

Figure 2-3: The newly discovered AI virus "GoldPickaxe"

Through the deeply forged fake face video, fraud molecules can log in the user's bank account, intercept the mobile phone verification message sent by the bank, and can transfer money, consume, change the account password, etc.

3 The "Deepfake" industrial chain

3.1 the "Deepfake" fraud process

With Dingxiang defense cloud business security intelligence center to intercept a "Deepfake" financial fraud cases, for example, fraud molecules fraud process mainly has four stages, "Deepfake" technology is just a key factor in the process of fraud link, other links, the victim if unable to identify and judgment, it is easy to fraud molecular instructions step by step into the trap.

The first stage, to cheat the victim's trust. Fraudsters contact the victim through text messages, social tools, social media, phone calls, etc., etc. (for example, can directly tell the name of the victim's name, home, phone, unit, address, ID number, colleagues or partners, or even some experiences) and gain trust.

The second stage, steal the victim's face. Fraudsters make video calls with victims through social networking tools, video conferences, video calls, and other methods. During the video call, they obtain the victim's face information (face, bow, turning, mouth, blink and other key information) to make "Deepfake" fake videos and portrait production. During this period, victims will also be induced to set up phone call transfer or induce victims to download malicious App software, which can call transfer or intercept the bank's customer service phone or phone.

The third stage, log in to the victim's bank account. The fraudsters log in the victim's bank account through the bank App, submit the fake videos and fake portraits made by "Deepfake", pass the bank's face recognition authentication, and intercept the mobile phone SMS verification code and risk tips sent by the bank on the victim's mobile phone.

The fourth stage, transfer the victim's bank balance. The bank calls the manual access phone call to the virtual number set by the fraudsters, posing as the victim, through the manual verification of the bank customer service staff, and finally successfully transferred the balance of the victim's bank card.

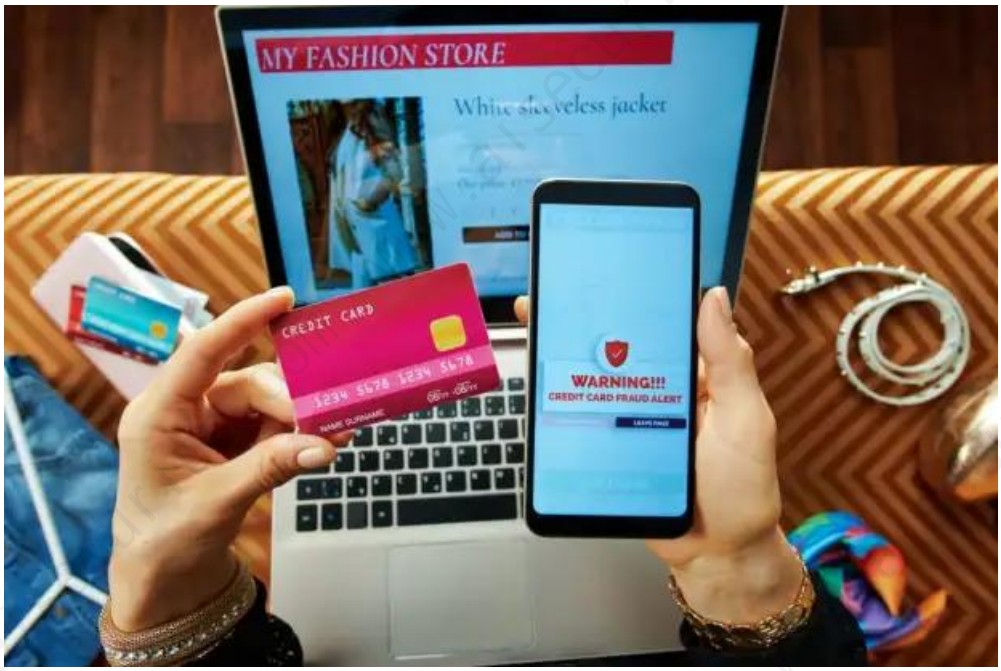

Figure 3-1: The ultimate purpose of the "Deepfake" fraud is to make a profit

"Deepfake" danger is not only to generate false video and images, but also contributed to the whole fraud ecosystem: a robot, false accounts and anonymous service of intricate network operation, all of these are designed to make, enlarge and distribute fabricated information and content, with difficult identification, detection and traceability problems.

Identification is difficult. Has developed to the point of convincingly generating realistic personal simulations, making it increasingly difficult to distinguish between true and false content to identify unless specifically trained, realizing that this threat is the first step in defending against it.

Detection is difficult. To improve the quality of "Deepfake", detection is a big problem. Not only the naked eye can not effectively identify, some conventional detection tools can not be found in time.

Traceability is difficult. There is no digital fingerprint, simulator forged address, false IP address, false device information, no clear digital clues to follow, or even no direct malicious software signature that can be detected, and traditional network security measures cannot effectively protect them.

It is a form of threat in the digital age, in which attackers are invisible and elusive, who not only produce information but also manipulate the realistic structure perceived by each participant. Therefore, to combat "Deepfake" fraud, on the one hand, it needs to identify and detect fake videos, pictures and information (effectively identify fake content); on the other hand, it is necessary to identify and detect "Deepfake" mode channels and platform networks (to improve the security of digital accounts in many aspects). This requires not only technical countermeasures, but also complex psychological warfare and the promotion of public safety

awareness.

3.2 "Deepfake" fraud upstream: the production of fake pictures and videos

The survey found that GitHub (software project hosting platform, the largest Web-based platform for hosting and managing collaboration and version projects) has more than 3,000 repositories related to "Deepfake" technology, indicating its broad development and distribution potential. On a foreign dark web tool, there are nearly a thousand channels or groups that provide "Deepfake". From fake video self-service production to personalized customization, everything. These "Deepfake" services are priced differently, with the lowest priced "Deepfake" videos running for $2 and a complex "Deepfake" video starting at $100, and ease of use making it easier for criminals to perform "Deepfake" fraud.

Figure 3-2: Industrial chain of the "Deepfake" fraud

With the advent of cybercrime as a service (Cybercrime as-a-Service), it is easy for ordinary people to buy Deepfake services or technology.

Dingxiang The Defense Cloud Business Security Intelligence Center analysis found that creating a "Deepfake" video involved the following process.

First step, collect the data. A large amount of data for target tasks is collected, including multi-angle face photos, work information, life information, etc., among which a large number of pictures, information and videos are collected on public social media.

The second step is the feature extraction. Deep learning algorithms are used to accurately identify and extract key facial features such as eyes, nose and mouth.

Figure 3-3: The production process of the "Deepfake" face change

The third step is portrait synthesis. Cover and fuse the visual faces to the faces that need to fake the tasks in the video, and align the facial features to ensure that they match and replace between the source and the target.

Step four, sound processing. Machine learning and artificial intelligence are used to replicate a person's voice, such as pitch, tone, and speech style, with amazing accuracy, and to match the lip movements in the video to synthetic speech.

Step 5: Render of the environment. Use lighting and color tools to further improve the coordination and matching of characters, voice, movements with the environment and clothing in the video.

Step 6, video synthesis.

3.3 "Deepfake" fraud downstream: multiple platforms and multiple ways to implement fraud

In 2021, the first domestic business security book "Attack and defend the road —— Enterprise digital business security risk and prevention" believes that the network fraud gang has formed a large-scale, organized group organization with clear division of labor for the purpose of

illegal profit, and spontaneously formed an ecological chain of the upper, middle and lower. In this black and gray production chain, there are groups specializing in making tools, group owners to provide all kinds of sensitive information, and groups to use tools and information to carry out all kinds of fraud. Black and ash production group has both bad users, and professional fraudsters, professional tools are constantly updated, new means emerge in an endless stream, it is difficult to fight against the defense.

Based on "Deepfake" fake audio and video and graphic information, fraud molecules using the simulator, machine software, IP seconds dial, registration, group control tools, will be able to bypass the bank, social media, remote video conference platform security detection, through remote meeting, work network, social platform and other different channels for various kinds of fraud.

Social media platforms have become the main distribution channels of "Deepfake" content, and finance, government affairs and enterprises are more important targets for fraudsters to carry out fraud. Dingxiang Defense cloud business security intelligence center analysis, fraudsters use "Deepfake" content for fraud, will use the following technical tools.

Figure 3-4: The scammers use various technical means

Seconds dial IP. IP address is the network information address when the Internet. This kind of tool can automatically call the global dynamic IP address, and can automatically switch, disconnection redial, automatically clean the browser Cookies cache, virtual network card information and other functions, can quickly and seamlessly switch the IP address of different areas.

imitator. GPS positioning is the geographical location information of users when using network services. Using simulation software and third-party tools, the user can change the longitude and latitude of the location where it is, and can realize the instant "crossing" anywhere.

Change machine tools. The model, series code, IMEI and so on of the device are unique. The device interface can be hijacked from the system level. When the application calls these interfaces to obtain the parameters of the device, the attribute information of the device that is forged by the modification tools is obtained. Generally speaking, 2~3 minutes to complete 1000 device attributes.

Registry machine. Registration is the key process for creating an account. The use of the registration machine can be batch automated account registration, so as to register hundreds or even tens of thousands of accounts.

group control. Can achieve a computer control on dozens, hundreds or even thousands of devices, unified registration, login, application. Group control also provides simulation positioning, shake, batch import address book and other functions, but also can be message push. The cloud control system of a company exposed by the "CCTV 315 Gala" in 2023 is this tool.

4 Identify the audio and video for the detection of the "Deepfake"

Enterprises and individuals need to verify their identity through multiple channels and adopt multiple strategies to identify and defend against "Deepfake" fraud. When everyone is likely to be targeted, while enjoying the convenience of technology, we should also agree to enhance our own protection awareness and ability. The best precautions are often not a technological solution or software, but a human factor. Therefore, the technology + human dual measures can be effective defense.

4.1 Human behavior and biometrics

4.1.1 During the video dialogue, you can ask the other party to press the nose and face to observe the face changes. If the nose is a real person, it will be deformed. You can also ask them to eat food, drink water, and observe the changes in the face. Or, ask for some strange movements or expressions, such as asking someone to wave or make a difficult gesture, to tell the truth. In the process of waving, it will cause the interference of facial data, which will produce a certain jitter or some flicker, or some abnormal situation. In the one-on-one communication, you can ask some questions that only the other person knows, to verify the authenticity of the other person. At the same time, when someone requests a remittance in a video or recording, they must call or verify repeatedly from other channels.

4.1.2 In the peer-to-peer communication, you can ask some questions that only the other party knows, so as to verify the authenticity of the other party.

4.1.3"Deepfake" can copy sound, but it may also contain unnatural intonation, rhythm or subtle distortion that will stand out after careful listening carefully. At the same time, speech analysis software can help identify speech abnormalities.

4.2 Equipment and operation identification

4.2.1 Digital signatures and blockchain ledger are unique and can track the source of behavior and mark them for review.

4.2.2 Compare and identify equipment information, geographical location and behavioral operations to find and prevent abnormal operations. Dingxiang Device Fingerprinting By recording and comparing the device's fingerprints, you can identify legitimate users and potential fraud behaviors. Its unique identification and identification technology, identify the virtual machine, agent server, simulator is malicious manipulation equipment, analysis equipment whether there is more account login, whether frequently change IP address, frequent change equipment properties such as abnormal or does not conform to the user habit behavior, help track and identify the activities of the fraud.

4.2.3 Remote account login, device replacement, changing mobile phone number, dormant account is suddenly active, etc. In addition, continuous authentication during session is critical and maintain persistent check to ensure that the user's identity is consistent during use. Dingxiang atbCAPTCHA It can quickly and accurately distinguish whether the operator is human or machine, accurately identify fraud behavior, and monitor and intercept abnormal behavior in real time.

4.2.4 In addition, restrict access to sensitive systems and accounts based on the minimum authority principle, ensuring access to the resources needed for their roles, thus reducing the potential impact of account theft.

Figure 4-1: Dingxiang Full-link panoramic face security threat perception scheme

4.2.5 Based on AI technology and artificial audit of face anti-flag, system, to prevent "Deepfake" fake videos and fake pictures. Dingxiang Full-link panoramic face security threat perception scheme, can effectively detect and identify the "Deepfake" false videos and false pictures. It for face recognition scene and key operation of real-time risk monitoring (such as camera hijacked, equipment, screen sharing, etc.), and then through the face environment monitoring, living recognition, image identification, intelligent verification multidimensional information check, found fake video or abnormal face information, can automatically block abnormal or fraud operation.

4.3 Technical identification and evidence collection

4.3.1 The generative adversarial network (GAN) based on deep learning can train a neural network model called "discriminator". Through training, the "discriminator" can more accurately identify the false image. The goal is to distinguish the real from the fake image video and identify any difference between the real version and the created version. Big data models can quickly analyze large amounts of video and audio data to identify anomalies at speeds beyond human capabilities. Moreover, the machine learning model can identify the characteristic patterns of the "Deepfake" production algorithm, thus identifying the content of the "Deepfake". And machine learning models can be retrained and tuned to maintain iterative evolution in real time.

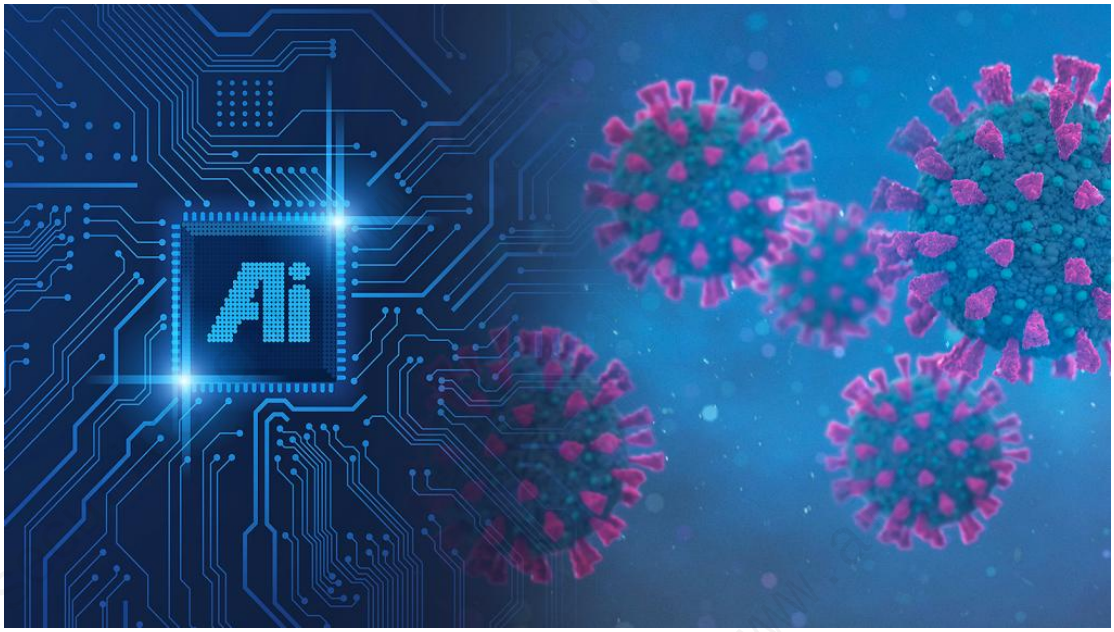

Figure 4-2: Use the AI against the "Deepfake" fraud

4.3.2 AI forensics tools play a vital role in investigating and attributing "Deepfake" content, analyzing digital footprints, metadata and other traces left in the creation process, helping to identify the attacker's sources and assist in legal investigations.

4.3.3 In case of document documents, the automatic document verification system can analyze document inconsistencies, such as font changes or layout differences.

4.4 Social prevention and public education

4.4.1 Reduce or eliminate the sharing of sensitive information such as accounts, families, transportation, and jobs on social media, and prevent fraud elements from stealing the "Deepfake" of pictures and sounds.

4.4.2 Continuous education to the public on "Deepfake" technology and its related risks is crucial, encouraging the public to remain vigilant and quickly report abnormal situations, and can also significantly improve the ability of organizations to detect and respond to "Deepfake" threats.

Technology is evolving, and new fraud are emerging. Keep abreast of the latest developments in AI and "Deepfake" technology to adjust safeguards accordingly. Continuous research, development and updating of AI models is essential to stay leadership in the increasingly complex "Deepfake" technology.

5 the security strategy against the "AI face change" fraud

5.1 Security strategies in social media scenarios

Fraudsters on social media platforms use registration machines to automate account registration on a large scale. In order to avoid the risk control and protection of social platforms, they also use machine modification tools to fake device attribute information to bypass the platform's security measures. Shockingly, it takes just two to three minutes to fake 1,000 device attributes, which allows fraud molecules to create a large number of fake accounts.

Figure 5-1: Security strategies in the social media scenario

In response to the problem of "Deepfake" fake information rampant on social media platforms, Dingxiang Defense Cloud Business Security Intelligence Center suggests that we should first accurately identify fraud accounts and then effectively prevent fraud behavior.

For the detection and identification of frame account registration process, it is suggested to timely find out whether there are risks such as swiping machine, Root, jailbreak, hijacking registration, risk injection, hook, simulator, etc. By comparing with the real machine, we can effectively distinguish the simulator cheating behavior, such as operator information, system information, hardware information, user behavior, CPU instructions, and simulator characteristics.

For the detection and identification of equipment used by fraud accounts, it is suggested to quickly identify the multiple activation of the same device, the abnormal IP behavior associated with the device, the large aggregation of IP in a short time, the abnormal proportion of old equipment models in the same channel, and the abnormal proportion of old operating systems in the same channel. At the same time, it is suggested to establish a dynamic operation and

maintenance mechanism of local list to precipitate and maintain the corresponding black and white list data through registration data, login data and activation data, including the blacklist of user ID, mobile phone number, equipment data and other dimensions.

For the monitoring and discovery of abnormal behavior of fraud account, it is suggested to conduct policy control based on user behavior. Through the risk control data and business precipitation data, the registration, login, browsing and release behaviors are modeled, so as to identify abnormal operations in time. The output of these models can be directly used in risk control strategies to improve the security of social media platforms.

5.2 Security strategy in the bank credit scenario

Loan fraud, theft brush, insurance fraud and other losses caused to the bank. In particular, the false audio and video made by using "Deepfake" technology stole other people's accounts and transferred the bank balance, which brought great economic losses to customers.

Figure 5-2: Security strategies in the Financial business scenario

Dingxiang The Defense Cloud Business Security Intelligence Center recommends that, in addition to strengthening customer authentication and authorization management, it also needs

to monitor abnormal operations in real time and identify potential risk accounts through AI.

First of all, through the use of multi-factor authentication, advanced authentication technology and other measures to improve the security of accounts, increase the reliability of authentication, to prevent accounts from being defrauded. Regularly update customer information to ensure the accuracy of the information and establish a complete customer profile, ensuring that only legitimate customers can access and use the account.

Secondly, the Dingxiang defense cloud and the Dingxiang Dinsight risk control engine are used to build a risk control system that combines machine learning and rule orientation. Through big data matching and tracking, multi-dimensional and in-depth analysis can accurately identify abnormal operations and carry out in-depth portrait of users, timely find suspicious operations, assess risks, and track and prevent various abnormal behaviors.

In addition, dig up hidden risk accounts. In order to avoid the audit of financial institutions and anti-fried mechanism, fraudsters will use various technical means to increase the difficulty of identifying abnormal behaviors, and the Dingxiang Xintell intelligent model platform can be used to explore the hidden and complex fraud behavior.

6 National Supervision and relevant laws and regulations

"Deepfake" is based on AI, and AI governance is crucial to the destiny of all mankind and is a common task facing all countries in the world. Many countries and organizations around the world have issued initiatives or norms, unanimously calling for strengthening the safety supervision of AI.

By imposing legal consequences on individuals who create or spread deep forgery, the spread of harmful content can be prevented and the perpetrators responsible for their actions. The "Deepfake" fraud qualitative criminal crime can also be a deterrent, preventing the abuse of the technology for fraud or other malicious purposes, is an effective way to mitigate the harmful effects of the "Deepfake" technology.
So countries need to address this urgent challenge and they need to implement strong legal measures. For example, criminal penalties must be imposed on those who intentionally create or promote harmful deep forgery transmission.

6.1 China

6.1.1 On July 13,2023, the Cyberspace Administration of China, together with relevant state departments, promulgated the Interim Measures for the Management of Generated Artificial Intelligence Services. Generative artificial intelligence services with the nature of public

opinion or the ability of social mobilization shall carry out security assessment in accordance with the relevant provisions of the State, and go through the procedures for algorithm filing, alteration and cancellation in accordance with the Provisions on the Management of Algorithms of Internet Information Service Algorithms.

6.1.2 In September 2023, the Ministry of Public Security of the Supreme People's Procuratorate of the Supreme People's Court issued the Guiding Opinions on Punishing cyber Violence crimes in accordance with the Law. Article 8 clearly stipulates that "using 'deep synthetic' generated artificial intelligence technologies and other generated technologies for publishing illegal information" will be given heavier punishment.

6.1.3 On October 18,2023, the Cyberspace Administration of the CPC Central Committee released the Global Artificial Intelligence Governance Initiative, whose specific measures include promoting the establishment of a risk level test and evaluation system, the implementation of agile governance, classified and hierarchical management, and rapid and effective response. Research and development entities need to improve the interpretability and predictability of AI, improve the authenticity and accuracy of data, ensure that AI is always under human control, and create audited, supervised, traceable and reliable AI technologies. At the same time, we will actively develop the development and application of related technologies for AI governance, and support the use of artificial intelligence technology to prevent risks and improve governance capacity. In addition, the initiative emphasizes the gradual establishment and improvement of laws and regulations to ensure personal privacy and data security in AI development and application, and opposes the illegal collection, theft, tampering and disclosure of personal information.

6.1.4 On March 1,2024, the National Technical Committee for Standardization of Network Security issued the Basic Requirements for the Security of generative ARTIFICIAL Intelligence Services, which clarified the basic security requirements of generative artificial intelligence services, including corpus safety, model safety and safety measures. Among them, in terms of corpus content security, service providers need to pay attention to three aspects: corpus content filtering, intellectual property rights and personal information. In terms of personal information, it is emphasized that before using the corpus containing sensitive personal information, the corresponding individual personal consent should be obtained separately or under other circumstances that comply with the provisions of laws and administrative regulations.

6.2 America

6.2.1 on October 30,2023, President of the United States Joe biden issued "safe, reliable and reliable artificial intelligence" administrative order, requires multiple government agencies standards, testing of artificial intelligence products, seek "watermark" validation, network security plan, attract technical personnel, to protect privacy, promote fair and civil rights, safeguard the interests of consumers and workers, promote innovation and competition, promote the leadership of the United States, etc. At the same time, the executive order protects U. S. users from AI fraud and deception by establishing standards for detecting AI-generated

content and authenticating official content.

6.2.2 On February 8,2024, the Federal Communications Commission has outlaw AI-generated sounds in robot calls, and lawmakers in various states have introduced legislation to combat false and false information generated by AI.

6.3 The European Union

6.3.1 In June 2023, the European Parliament passed the draft authorization of the EU Artificial Intelligence Act. The bill divides AI systems into four categories based on risk levels, from minimum to unacceptable. Among them, "technology robustness and security" requires the AI system to minimize accidental injuries in the process of development and use, and to have the robust ability to deal with unexpected problems, so as to prevent malicious third parties from illegally using the system or changing its way of use or performance. In addition, the bill bans the establishment or expansion of facial recognition databases through untargeted extraction of facial images from the Internet or CCTV videos, and bans the use of such methods to put AI systems on the market, into use or into use. For generative AI systems based on these models, the bill requires compliance with transparency requirements, meaning that disclosure is generated by AI systems, and ensures that illegal content is not generated. Furthermore, a detailed summary of these data must be disclosed when using copyrighted training data. For example, article 52 (3) requires creators to be transparent. This means that anyone who creates or spreads "Deepfake" must disclose its artificial sources and provide information about the technology used.

6.3.2 On February 21,2024, the establishment of the EU AI Office, marking an important step in promoting responsible AI practice within the EU. One of the main functions of the AI Office is to encourage and facilitate the development of codes of conduct at the ITU level to facilitate the effective fulfillment of obligations related to the detection and labeling of manually generated or manipulated content. Under this assignment, the Commission has the authority to approve these codes of practice by enforcing the Act. This regulatory mechanism ensures that the code of conduct meets certain standards and effectively responds to the challenges posed by artificially generated or manipulated content. Moreover, if the Commission considers that a certain code of practice is inadequate, it has the power to address any defect by implementing the Act.

6.4 Australia

6.4.1 In July 2023, the ACCC and the Australian Securities and Investment Commission (ASIC) established the Investment Fraud Integration Group of the National Anti-Fraud Centre to combat investment fraud. According to Lowe, investment fraud losses reported to Scamwatch have declined since the group was formed.

6.4.2 Australian authorities on Friday warned their people to be alert to misinformation

after the National Anti-Fraud Center revealed that 400 people reported losses of more than 8 million Australian dollars (US $5.2 million) in 2023 due to fraud on online trading platforms.

Katrona Lowe (Catriona Lowe), vice chairman of the Australian Competition and Consumer Council (ACCC), said the scams operate by creating fake news articles and videos featuring celebrities."Fraudsters are making fake news articles and deeply fake videos to convince people that celebrities and prominent public figures are using online investment trading platforms to make huge amounts of money when this is actually a scam.”

6.5 International organizations and conferences

6.5.1 On November 1,2023, at the first Global Artificial Intelligence (AI) Security Summit, 28 countries signed the Bletchley Declaration on international governance of AI, which is the world's first international statement on artificial intelligence, a rapidly emerging technology. The Declaration encourages relevant actors to take appropriate measures, such as safety tests, assessments, to measure, monitor and mitigate the potential harm of AI and its possible effects, and to provide transparency and accountability.

6.5.2 On October 30,2023, the Group of 7 (G7) issued the International Code of Conduct for Developing Advanced Artificial Intelligence Systems Organizations. The code of conduct contains 11 items and highlights the measures to be taken during development to ensure credibility, security, and security. Among them, developers need to identify and mitigate risks, including Red Team testing, testing, and mitigation measures. At the same time, developers also need to identify and reduce vulnerabilities, events, and misuse patterns after deployment, including monitoring vulnerabilities and events, and facilitate third parties and users to detect and report problems. Furthermore, the guidelines highlight the importance of developing and deploying reliable content authentication and source mechanisms, such as watermarking. These measures will help ensure the security and reliability of AI systems and improve users' trust in them.

6.5.3 At a security conference in Munich, Germany on February 17,2024,20 world leading technology companies including AWS, Google, IBM, LinkedIn, McAfee, Meta, Microsoft, OpenAI, Snap, TikTok and X announced that they will jointly combat "Deepfake" information. At the meeting, they collectively signed a technical agreement to resist deceptive AI-generated content, reduce the generation of deceptive AI content and its risks, and agreed to propose solutions on their respective platforms or products. The deal also promises to work with organizations and academia around the world to make the public and the media aware of the dangers of deceptive content generated by AI.

7 Which institutions should take responsibility

"Deepfake" fraud needs to be responsible?

7.1 Social media

Platforms often position themselves as mere content channels, but such protection does not exist in several countries, including Australia. The Australian Competition and Consumer Commission (ACCC) is taking the Facebook to court. The ACCC also argued that Facebook should be held liable as an accomplice to the scam —— for failing to remove misleading ads in time after receiving notice of the issue. At the very least, the platform should be responsible for timely removing deep forgery for fraud purposes. Many platforms already claim to be doing so, for example, on Facebook, any AI-generated content may display an icon that clearly shows that the content is generated with artificial intelligence.

Social media companies have the most influential in limiting the spread of fake content, being able to detect and remove it from their platforms. However, the policies of the major platforms, including Facebook, YouTube and TikTok, stipulate that they will only remove Midea content if they "seriously hurt" or aim to mislead people during the voting process. This coincides with a general relaxation of audit standards, including the abolition last year of 17 policies for the first three companies related to hate speech, harassment and misinformation.

7.2 Financial institutions

Britain is introducing a mandatory scheme requiring banks to compensate, at least in some cases, for victims of fraud. In other words, the bank is required to compensate for the loss of users in the "Deepfake" scam. Currently, the ACCC and others in Australia have proposed similar plans, but not at this stage.

7.3 AI tool provider

Providers of AI tools are currently not legally supervised or control whose tools are not being used for fraud, but AI tool provider enterprises do have the opportunity to use technology to reduce the possibility of "Deepfake". As with banks and social media platforms, AI tool providers may soon be asked to do so. For example, the recently introduced EU Artificial Intelligence Act requires providers of born AI tools to design these tools in a way that allows the detection of synthetic / fake content, add digital watermarking, timely limit digital identity cards used to verify individual identity, etc.

AI watermarking works by embedding unique signals into the output of an AI model, which can be an image or text designed to identify content as content generated by AI to help others identify it effectively.

URL:www.aisecurius.com
Email address: marketing@dingxiang-inc.com