Issues in Information Systems
Volume 25, Issue 2, pp. 222-229, 2024

DOI: https://doi.org/10.48009/2_iis_2024_118

Artificial intelligence in cyber threats intelligence (CTI): capabilities of current AI models for deepfakes

Mohammed Almuthaybiri, *Marymount University, mma84343@marymount.edu*
Gaston Elongha, *Marymount University, gle85144@marymount.edu*
Innocent Mgembe, *Marymount University, ipm87755@marymount.edu*

Abstract

Artificial Intelligence's advancement has led to ethical and privacy concerns due to the ability of algorithms to mine personal data and conduct surveillance on a large scale. The influence of Artificial Intelligence (AI) in cybersecurity extends beyond private entities and individuals, but also governments, and local and foreign nations. Due to this challenge, it is essential to highlight the threatening actors and their objectives to counter AI-based cyber threats like deepfake threats. Cyberspace deals with different intentional and non-intentional entities, each having its interests: cybercriminals are interested in making a profit, terrorist groups cause ideological violence, and nation-states target geopolitical influence. The emergence of a new generation of digitally manipulated media called "deepfakes" is not just a technological marvel but a potential threat that is revolutionizing cyber actors' operations. Deepfakes as a threat can generate highly realistic and compelling videos, which have raised substantial concerns about their usage's possibility .

Keywords: Artificial Intelligence, cyber threat intelligence, deepfake, threat actor, emerging technology

Introduction

During the last decade, social media platforms developed to connect people and help share ideas and opinions through multimedia content (images, video, audio, and texts) are nowadays being used to manipulate and alter public opinion. Thanks to bots, i.e., computer programs that control a fake social media account as a legitimate human user would do by "liking," sharing, and posting old or new media which could be genuine, forged through simple techniques (e.g., editing of a video, use of gap-filling texts and search-and-replace methods) or deep fake.

Deep learning is a family of machine-learning methods that use artificial neural networks to learn a hierarchy of representations of the input data from low to high non-linear features representation. Deepfake is a portmanteau of "deep learning" and "fake" (Fagni, et al., 2021). Five years ago, nobody had even heard of deepfakes, the persuasive-looking but false video and audio files made with artificial intelligence. (Metz, 2022).

While AI has undoubtedly enhanced cybersecurity, it has also inadvertently given hackers an advantage in executing sophisticated attacks. The emergence of AI-based deepfake technology has further expanded the possibilities for good and bad actors. Deepfake technology relies heavily on machine learning, enabling it to create convincing but false media at a faster pace and lower cost. This paper will explore the potential

capabilities an attacker could develop using AI technology, including deepfakes (Petru-Dan, 2023).

Background & Related Works

Deepfake technologies first emerged in computer vision, followed by effective audio manipulation and text generation attempts. Recently, audio deepfakes involved generating speech audio from a text corpus by using the voices of different speakers after five seconds of listening time (Fagni, et al., 2021).

Deepfakes are frequently utilized for exploitation despite their harmless and legal uses in industries like entertainment and commerce. Many recent authors and believers claim that most deepfake recordings target politicians or celebrity personalities, arguing that they are widely disseminated online for misinformation.

Deepfakes could be employed as a psychological warfare tool to sway elections or stir up civil unrest. They may also cause people to ignore valid proof of misconduct and, in general, erode public confidence in audiovisual content (Polovinkin & Low, 2024).

With that in mind, the research's main question was: *How does deepfake technology impact geopolitical interests and personal identifications?*

Fake images and videos are the most common payload in the deepfake domain. This could be due to their further appeal to human audiences (Azmoodeh & Dehghantanha, 2022). The underlying technology can synthesize speech, replace faces, and control facial emotions. In a deepfake, someone can appear to speak or do something that they never said or did. Face swapping and facial expression manipulation are usually prominent in deepfake videos. Academics and internet firms have tested several techniques to identify deepfakes. These techniques typically employ AI to scan videos or photos for digital flaws or details—like blinking or facial tics—that deepfakes cannot replicate accurately (Petru-Dan, 2023).

Threat actors already use deepfake content to impersonate executives and political leaders in targeted social engineering attacks. Threat actors can effectively impersonate key leadership figures and connect audio and video deepfake content to conference calls or VOIP technology using publicly available content such as videos, interviews, and pictures.

Insikt Group researchers generated deepfake videos impersonating four Recorded Future executives promoting a fictional sponsorship deal. Using open-source video interviews and internal conference call recordings, we could train Tortoise TTS, an open-source voice impersonation tool, to impersonate these executives without demonstrating their consent and overlay the resulting audio over existing conference call footage (InsiktGroup, 2024b).

In early 2020, a United Arab Emirates bank was tricked by cybercriminals using AI voice cloning to steal $35 million. According to court documents recently uncovered by Forbes, the fraudsters used the deep-faked voice of a company executive to fool a bank manager into transferring millions of currencies to their possession.

The cyber threat actors employed AI voice cloning technology to create a convincing fake voice of the executive. They managed to get away with the stolen money, leaving the bank and the company to deal with the aftermath (Alvareztg, 2023; InsiktGroup, 2024a). ElevenLabs also offers paid users the option to upload “custom voices,” which allow public figures to be impersonated. This report will reference this option, but Insikt Group analysts refrained from generating or downloading such samples due to the high potential for abuse (InsiktGroup, 2023).

Figure 1: Banking Fraud (Source: ElevenLabs)

Figure 1 shows a sample prompt created with the premade voice "Adam" in ElevenLabs. Stability, clarity, and similarity enhancement were increased to create a more natural-sounding voice with artifacts instead of a monotone or robotic sample for Banking Fraud, and among these new fraud methods is an infrequent occurrence of a new sophisticated mobile Trojan, GoldPickaxe, aimed explicitly at iOS users.iOS by Group-IB targets users of over 50 Vietnamese banking applications.

GoldPickaxe.iOS, Group-IB researchers found that cyber threats actors employing deepfakes as their attack vector, can collect facial recognition data and identity documents and intercept SMS. Furthermore, Android has the same functionality but exhibits other functionalities typical of Android Trojans. The threat actor utilizes AI-driven face-swapping services to exploit the stolen biometric data to create deepfakes. The data stolen and deepfakes created, combined with identification documents and the ability to intercept SMS, enables cybercriminals to gain unauthorized access to the victim's banking account. This is a new technique of monetary theft previously unseen by (Group-IB) researchers (Polovinkin & Low, 2024).

Moreover, sophisticated deep fake technology trains programs to generate realistic impersonations of another's body and create a realistic image or video (*The porn industry*). Advancements have enabled users to input a written prompt, making a convincing, falsified image or video. Startups are working on developing detection technology, and the company with the text-to-image tool claims to have closed the loophole that allowed for the Taylor Swift deepfakes (Brenner & Ruiz-Lichter, 2024).

Malicious actors on the internet have been targeting teen girls and creating AI-generated deepfake images at unprecedented rates. This year, for example, New Jersey high schooler Francesca Mani spoke at a news conference alongside Congressman Joe Morelle, and she explained that nonconsensual Al-generated intimate images of her, along with the photos of 30 other girls at her school, were shared on the internet. An MIT Technology Review report revealed that the vast majority of deepfakes target women.

Between 90% and 95% of these videos are nonconsensual porn involving women (Brenner & Ruiz-Lichter, 2024). LastPass, a password management platform, recently disclosed an attempted voice phishing attack targeting one of their employees using sophisticated deepfake audio technology to impersonate the company's CEO, Karim Toubba. The threat actor attempted to contact the employee via WhatsApp, leaving a series of calls, text messages, and at least one voicemail that featured audio deepfake impersonation, which is an uncommon business channel used by the company and is originally why the employee identified the activity as suspicious (Ankura, 2024).

Such deepfake videos and fake news articles are increasing at unusual way and show the necessity to develop detection and mitigation approaches for such threats. Some believe that effective deepfakes

mitigation is to enhance the resilience of discriminative models and building methods for authenticating the produced contents (Mohamed et al., 2023).

Nation States - Geopolitical Interest

Deepfakes may not have much impact beyond drawing interest and getting traction online. Nevertheless, the negative impact increases in critical circumstances, for example, during war or natural disaster, when people cannot reason and only concise span of attention (Metz, 2022). This exacerbates when it comes to government officials and organizations via state sponsorship. Nation-states use AI to produce more credible misinformation campaigns to erode public confidence in democratic processes, social cohesiveness, and government institutions.

Also, terrorist groups sponsored by a state can create more intricate and sophisticated attacks because they have plenty of resources and experience. They may target critical infrastructure and industries within a nation, manipulate polls, and spread misinformation to undermine its constitutional framework or steal confidential data from businesses and government agencies. According to Homeland Threat Assessment 2024, Russia, China, and Iran are still developing the most advanced malicious influence operations online, and part of the operation figures deepfakes. These adversaries will employ many of the same strategies to sway US audiences in the run-up to the 2024 election. For instance, AI-generated deepfake technology can be used to sway (Petru-Dan, 2023).

Some examples that illustrate the potential power of deepfakes, especially the most convincing and potentially harmful deepfakes include:

- ***Former President of the United States Donald Trump (2018).*** This was an example of a politically motivated deepfake video encouraging Belgium to withdraw from the Paris climate agreement. The video was published by the Belgian Social Democratic Party to start a public debate regarding the necessity of addressing climate change.
- ***Former President of the United States, Obama (2018).*** The video was created using FakeApp and involved taking an original video of Barack Obama and pasting Jordan Peele's mouth into it (ThinkTank, 2021).
- ***CEO of Meta Mark Zuckerberg (2019).*** A video appeared on Instagram falsely portraying Facebook's CEO Mark Zuckerberg and crediting a secretive organization for the company's success. The video was created by Canny AI, an Israeli company, and constructed using deepfake technology, an actor's voice, and 2017 footage of Zuckerberg (Metz, 2019).
- ***The dispute between Saudi Arabia*** **and** ***Russia regarding oil production*** quotas could negatively impact the price of oil and, thus, the global economy. Bateman concludes saying that there is no serious threat to the stability of the global financial system or national markets in nature and healthy economies. However, manipulating news media, in turn, can damage economic stability (Metz, 2019).
- ***A bank manager in Hong Kong (2020)*** was defrauded into transferring thirty thousand-five million US dollars through a voice he recognized as the director of a company based in the United Arab Emirates. The deepfake call was supplemented with emails to the bank manager from the director confirming the transfer (Mannie, 2024).
- ***Former President of Gabon Ali Bongo Ondiba (2019).*** A video with Ali Bongo Ondiba was published online for months before even being seen in public, and that had become a widespread belief that he was in poor health, or even dead (Metz, 2019). The video led to a national crisis, a story that the video was a deepfake gained momentum, as it supported a theory that the government was trying to hide the actual and real condition of the president (Metz, 2019).

- ***Russian opposition politician Leonid Volkov (2021),*** and various news outlets reported several European Parliament members were targeted by deepfake video calls imitating Russian opposition; the calls were created by Russian pranksters (ThinkTank, 2021).
- ***President of Ukraine, Volodymr Zelensky (2022).*** The deepfake appeared on the hacked website of the Ukrainian TV network "Ukrayina 24," which showed Ukraine's President talking about surrendering to Russia. Volodymyr Zelensky appears behind a podium, telling Ukrainians to surrender their weapons (Allyn, 2022).
- ***President of Russia Vladimir Putin in 2022.*** A deepfake video shared on X shows Russian President Vladimir Putin declaring peace (Evon, 2022).

Methodology

The previous section of the paper strategically and extensively reviewed existing literature, studies, and case examples that illustrate deepfakes and how they are leveraged by cyber threat actors. The methodology employed in this research is a qualitative approach that focuses on understanding the nature, characteristics, and implications of AI in cybersecurity and deepfakes. That will be accomplished through detailed descriptions and analysis. Not only that, but the methodology also categorizes distinct types of cyber threat actors, their motivations, and methods, which are typical of qualitative research, aiming to understand phenomena in terms of their meanings.

The article delves into AI's ethical and societal implications in cybersecurity, with the aim of understanding human experiences and societal impacts (Petru-Dan, 2023). The researchers also employed an inductive, mixed methods research design (Creswell & Creswell, 2018), combining qualitative and quantitative analysis, using cybersecurity threat open-source platforms, and analyzing data using Recorded Future (RF) and Insikt Group's recent studies.

Because deceptive synthetic media could inflict financial harm on many potential targets, apparent targets include person, organization, and country identifications. Directly comparing synthetic media and open-source platforms will help illuminate the most concerning scenarios, especially those that could empower bad actors and require new policy responses. In addition, that helps identify the least concerning scenarios that are no more dangerous than today's threats and may not need further responses (Bateman, 2020).

Findings & Discussion

With recent alerts from the U.S. Department of Health and Human Services (HHS) on similar social engineering tactics and Al voice cloning tools that could be used to target IT help desks, the FBI also warned about the potential widespread use of deepfakes in cyber and foreign influence operations (Ankura, 2024). Snuffing out misinformation, in general, has become more complex during the war in Ukraine. Russia's invasion of the country has been accompanied by a real-time deluge of information hitting social platforms like X, Facebook, Instagram, and TikTok. Nevertheless, opinion is widely divided between the real and true from fake and misleading contents (Metz, 2022).

Fully reliable, and scalable defenses against voice and video deepfakes do not exist yet. Although, there are some measures organizations can take to make it more difficult for cyber criminals to succeed. One way to do that is to ensure awareness of deepfakes by ensuring employees and customers are educated about how convincing a deepfake imposter or content can be. Another way is through a multilayered identity verification process that requires entities to adapt processes involving financial transactions or sensitive data transfer to include multiple verification steps. A critical step includes requiring confirmation on a separate communication channel or limiting the ability to process high-risk transactions to a few highly

trained individuals. Finally, monitoring online information for any publicly available data regarding companies or leaderships, including images and videos of executives, as they can easily be weaponized as part of a deepfake fraud (InsiktGroup, 2024a).

Figure 2 shows all deepfake threat events reported and raised in the Recorded Future platform in the last two months we managed to extract and analyze. It shows how deepfakes are being mentioned and applied in many channels, not just the one previous researcher mentioned but other like telegram. With this graphic, we can see how deepfake is at the top of most even mentions about content (images, videos, texts). These findings respond to our research question and are supported by FBI (2022) Gatlan (2022) warnings around how deepfakes capabilities can be leveraged for cyber fraud, not just financial but employments, etc.

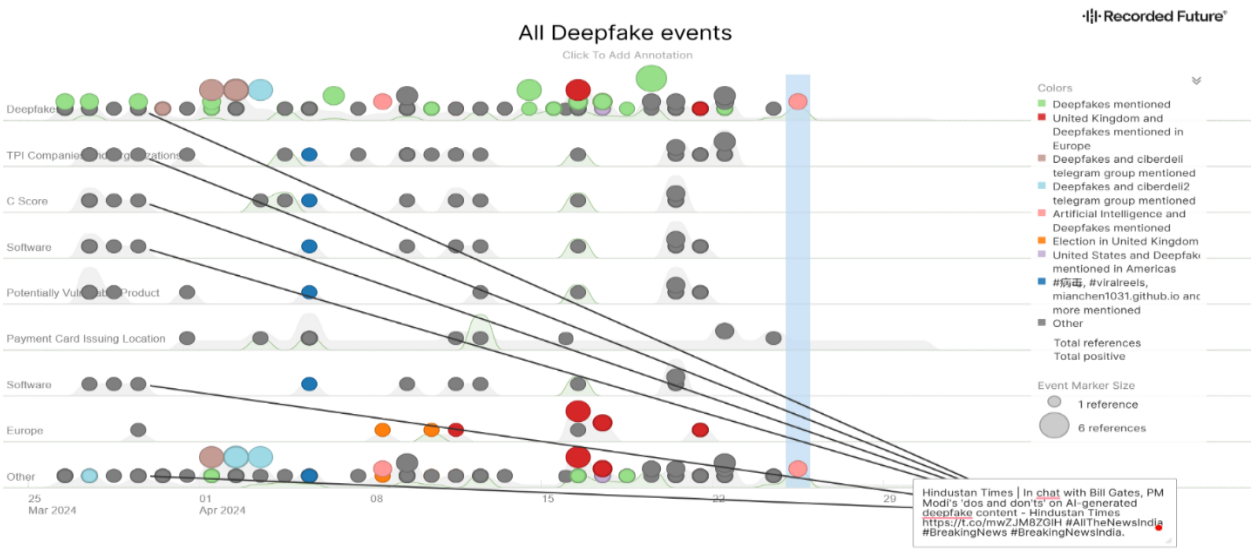

Figure 2: Deepfake events (Source: Recorded Future)

The FBI Internet Crime Complaint Center (IC3) warned of increased complaints reporting the use of deepfakes and stolen Personally Identifiable Information (PII) to apply for various remote work and work-at-home positions. In these interviews, the actions and lip movements of the person seen interviewed on-camera do not coordinate with the audio of the person speaking. At times, actions such as coughing, sneezing, or other auditory actions are not aligned with what is presented visually (FBI, 2022; Gatlan, 2022).

Research Implications & Conclusion

Integrating AI across various sectors, including cybersecurity and social networks, has positive and negative implications. This duality presents a critical challenge in the field. Deepfake technology, a prominent application of AI, poses significant risks. It primarily manipulates images and videos to create authentic but false representations of the actual subject. Although it has legitimate applications in entertainment and commerce, deepfakes are exploited for malicious purposes such as spreading misinformation, swaying public opinion, and undermining trust in audiovisual content. Nation-states use AI for geopolitical interests, including misinformation campaigns to undermine democratic processes and public trust. The complexity of the cyber threat landscape is increasing, with various actors employing AI to enhance their offensive and defensive capabilities. This, therefore, requires a balanced approach, combining AI's computational power with human expertise and judgment for effective cybersecurity (Petru-Dan, 2023).

The rise of deepfake technology also requires enhanced awareness and hardened security measures, including strict verification processes for sensitive requests, training for staff to recognize social engineering, and the revalidation of users with access to critical systems (Ankura, 2024). Synthetic media is unlikely to threaten global financial stability or the macro economy—but individuals, companies, and government entities are vulnerable, and so are emerging economies and their systems under stress (Bateman, 2020). The widespread availability of sophisticated deepfakes will fundamentally shift how we manage identity and establish trust in all digital mediums. Companies must balance considerations for ease of use and customer experience against securing their sensitive financial and personal data. These companies introduce "*one-to-many*" opportunities for threat actors, where compromising one business process outsourcing can provide data access to multiple client companies, making them increasingly appealing targets for sophisticated attacks (InsiktGroup, 2024a).

References

Alvareztg. (2023). Cybercriminals Utilizing AI Acquire Large Sum of Money from UAE Bank. Alvareztg. Retrieved May 05, 2024, from https://www.alvareztg.com/uae-bank-deepfake

Allyn, B. (2022). Deepfake video of Zelenskyy could be 'tip of the iceberg' in info War, experts warn. KUOW. https://www.kuow.org/stories/deepfake-video-of-zelenskyy-could-be-tip-of-the-iceberg-in-info-war-experts-warn

Ankura. (2024). Ankura CTIX FLASH Update. Ankura. Retrieved May 21, 2024, from https://angle.ankura.com/post/102j5gf/ankura-ctix-flash-update-april-16-2024

Azmoodeh, A., Dehghantanha, A. (2022). DEEP FAKE DETECTION, DETERRENCE AND RESPONSE: CHALLENGES AND OPPORTUNITIES. arXiv.org. https://arxiv.org/pdf/2211.14667

Bateman, J. (2020). Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios. Carnegie Endowment for International Peace. Retrieved May 02, 2024, from https://carnegie-production-assets.s3.amazonaws.com/static/files/Bateman_FinCyber_Deepfakes_final.pdf

Brenner, N., & Ruiz-Lichter, S. (2024, April 22). Why the Taylor Swift AI Scandal is Pushing Lawmakers to Address Pornographic Deepfakes. Lexology. https://www.lexology.com/library/detail.aspx?g=118491c5-d761-4cd1-b33e-83e08c351839

Creswell,J. D. Creswell, J. W. (2018). Research design: Qualitative, quantitative, and mixed methods approach. SAGE Publications. Retrieved May 14, 2024, from https://us.sagepub.com/en-us/nam/research-design/book270550

Evon, D. (2022). Putin Deepfake Imagines Russian President Announcing Surrender. Snopes. Retrieved May 03, 2024, from https://www.snopes.com/fact-check/putin-deepfake-russian-surrender/

Fagni, T., Falchi, F. Gambini,M. Martella, A. Tesconi1, M. (2021). TweepFake: about detecting deepfake tweets. arXiv.org. https://arxiv.org/pdf/2008.00036

FBI. (2022). Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions. Federal Bureau of Investigation [Report]. https://www.ic3.gov/Media/Y2022/PSA220628

Gatlan, S. (2022, June 28). FBI: Stolen PII and deepfakes used to apply for remote tech jobs. BleepingComputer. https://www.bleepingcomputer.com/news/security/fbi-stolen-pii-and-deepfakes-used-to-apply-for-remote-tech-jobs/

Insikt Group. (2023). I Have No Mouth, and I Must Do Crime. Retrieved May 03, 2024, from https://www.recordedfuture.com/i-have-no-mouth-and-i-must-do-crime

InsiktGroup. (2024a). Adversarial Intelligence: Red Teaming Malicious Use Cases for AI. Retrieved May 03, 2024, from https://www.recordedfuture.com/adversarial-intelligence-red-teaming-malicious-use-cases-ai

Insikt Group. (2024b). Preparing Your Company for the Coming Era of Deepfake Deception. Retrieved May 01, 2024, from https://www.recordedfuture.com/search-results?&query=deepfakes&page=1

Mannie, K. (2024). Company out $35M after scammers stage video call with deepfake CFO, coworkers. Global News. Retrieved May 22, 2024, from https://globalnews.ca/news/10273167/deepfake-scam-cfo-coworkers-video-call-hong-kong-ai/

Metz, R. (2022). Deepfakes are now trying to change the course of war. CNN Business. Retrieved May 12, 2024, from https://amp.cnn.com/cnn/2022/03/25/tech/deepfakes-disinformation-war/index.html

Mohamed, A., Merouane, D., Muna, Al-Hawawreh. (2023). Generative AI for Cyber Threat-Hunting in 6G-enabled IoT Networks. arXiv.org. https://arxiv.org/pdf/2303.11751

Petru-Dan, K. (2023). THREAT ACTORS SEEKING TO EXPLOIT AI CAPABILITIES. TYPES AND THEIR GOALS. National Defense University [Journal-article]. Retrieved May 13, 2024, from https://pdfs.semanticscholar.org/17be/6c656c31b476b99ea0ede2544bcb10d6a6ce.pdf

Polovinkin, A., Low, S. (2024). Face Off: Group-IB identifies the first iOS trojan stealing facial recognition data. Retrieved May 02, 2024, from https://www.group-ib.com/blog/goldfactory-ios-trojan/

Think Tank. (2021). Tackling deepfakes in European policy. European Parliament. Retrieved May 07, 2024, from https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039+