Understanding the Risk: Mapping Deepfake Cyberattacks to a Temporal Attack Model

Heloise Pieterse
National Integrated Cyberinfrastructure System
Council for Scientific and Industrial Research
Pretoria, South Africa
https://orcid.org/0000-0002-2908-4012

***Abstract*—The advancement of artificial intelligence (AI) technologies has become a trending topic in the cybersecurity domain. These technologies, however, present cybersecurity with a double-edged sword as AI offers enhanced threat detection and protection, but also enables cybercriminals to craft sophisticated cyberattacks. Deepfakes, which are a form of digitally manipulated synthesised media created using deep learning techniques, have garnered widespread attention due to the use of deepfakes in cyberattacks to cause influence, spread disinformation, or conduct fraudulent activities. While extensive research efforts have been undertaken to develop defences against deepfakes, the solutions are technical and not easily accessible. Innovative strategies are required to equip personnel from government, academia, and the business sector with the fundamental knowledge to detect and defend against cyberattacks employing deepfake technology. This paper evaluates the most significant events involving deepfakes since the emergence of the technology in November 2017. Key trends and characteristics are identified and mapped to a temporal attack model to separate the different stages of a cyberattack involving deepfakes. The outcome is a Deepfake Attack Framework that offers valuable insights essential to understanding the risks associated with deepfakes. The Deepfake Attack Framework presents a theoretical solution that can be applied practically to minimise risk and enable personnel to be better prepared to defend against deepfake-driven cyberattacks.**

Keywords—Artificial Intelligence, Deepfakes, Deep Learning, Attack Model, Cyberattacks, Cybersecurity Awareness.

I. INTRODUCTION

The Fourth Industrial Revolution (4IR) is being defined by the rapid technological advancement of the 21st century, driven by the convergence of technologies such as cloud computing, Internet of Things (IoT), blockchain, robotics, big data and Artificial Intelligence (AI). In particular, AI appears to be at the core of the revolution, accelerating the impact of 4IR on various sectors. The healthcare sector has witnessed a radical transformation with the adoption of AI, such as improved patient care, personalised diagnosis, and advanced drug discovery [1, 2]. Similarly, AI-driven automation improves the efficiency of manufacturing operations and streamlines product design, leading to optimised production [2, 3]. It is, however, the cybersecurity domain that is currently at the forefront of the AI revolution, benefiting from improved threat hunting, quick detection of anomalies, and predictive analysis. Although advancing cybersecurity, AI has become a double-edged sword for the domain [4]. As AI strengthens and improves cyber defences, cybercriminals leverage AI for more sophisticated attacks.

The term "deepfake" is an unification of "deep learning" and "fake" [5, 6], describing the creation of AI-generated "synthetic media" [5]. The synthetic media, usually videos, images, or audio, are created using deep learning algorithms, often Generative Adversarial Networks (GANs), to generate realistic content representing real people [6]. The proliferation of deepfakes is driven by two factors: technological advances and societal context [5]. Technological advances, such as improved computing power, sophisticated AI algorithms, and accessibility to pre-trained and developed tools, continue to enhance the quality of deepfakes [5, 7]. From a societal perspective, the dependence often placed on social media provides a fertile ground for the rapid and widespread dissemination of deepfakes, which can have serious consequences for individuals, organisations and the larger society [8].

The third annual Identity Fraud Report, released by Sumsub in 2023 [9], revealed a significant increase in deepfake cyberattacks detected globally. Although deepfake cyberattacks impact multiple sectors, the financial services industry is the hardest hit, with 88% of all detected deepfake attacks targeting the cryptocurrency sector, followed by 8% affecting the financial technology sector. More startling is the 1200% increase in deepfake fraud and cyberattacks discovered in South Africa in 2023 [10]. Deepfakes continued to surge in 2024 [11] with the significant increases observed in Africa (393%). The growing use of deepfakes poses a serious risk, fuelling financial fraud and potentially inciting political instability [12].

Extensive research efforts have been undertaken to develop defences against deepfake cyberattacks [13, 14]. Such defences often rely on deep learning techniques to detect irregularities, which include pixel inconsistencies, unnatural facial movements, abnormal blinking, or mismatched shadows. While technological solutions based on these defences have been developed, such solutions are highly technical and often inaccessible to the public [15]. Therefore, innovative strategies are required to equip government, academia, and the business sector personnel with the fundamental knowledge to detect and defend against cyberattacks employing deepfake technology.

Following a literature survey research methodology, this paper evaluates the most significant events involving deepfakes since the technology emerged in November 2017. The objective of this literature survey is to identify key trends and characteristics, which are then mapped to a temporal attack model to separate the different stages of a cyberattack involving deepfakes. The outcome is a Deepfake Attack Framework that offers a theoretical solution to support

education and cyber awareness campaigns to prepare the public to better defend against deepfake-driven cyberattacks.

The remainder of the paper is structured as follows. Section II presents the background on the evolution of deepfakes and introduces the temporal attack model. In Section III, the most significant deepfake cyberattacks are captured and discussed. The Deepfake Attack Framework is presented in Section IV and further discussed in Section V. The paper concludes in Section VI.

II. Background

The current proliferation of deepfakes necessitates the development of a Deepfake Attack Framework to provide a structured way of describing and presenting deepfake cyberattacks. This framework is grounded in a comprehensive understanding of deepfakes and is organised based on a temporal attack model.

A. Evolution of Deepfakes

Deepfakes have witnessed unprecedented growth in recent years, mostly due to technological advancements. While deepfakes remain a novelty to most individuals, manipulation of media is not a new trend. In 1860, as part of a propaganda campaign, a portrait of politician John Calhoun was cleverly manipulated by replacing his head with that of the current American President [16]. Another example is Video Rewrite, a system published in the late 1990s that used existing footage to automatically create a new video of a person mouthing words not spoken in the original footage [17].

However, with the introduction of the Generative Adversarial Network (GAN) in 2014 [18], a new era of manipulated multimedia was established. GANs enable the generation of highly realistic synthetic data by training two neural networks, a generator and a discriminator (see Fig. 1). Following an adversarial training process, the generator produces synthetic data, attempting to deceive the discriminator, while the discriminator learns to distinguish between real and fake data. As the training progresses, the output produced by the generator becomes increasingly more realistic, causing the discriminator to start struggling to differentiate between the generated and real data [20]. The final result is the production of realistic data.

The term "deepfake" first appeared in 2017 and is credited to an anonymous user, u/deepfakes, who posted the first deepfake video on Reddit, a popular social media platform of small communities focused on content sharing and topic-specific discussions. With the discovery of the first deepfake video, the community was named r/deepfakes, and an assortment of face-swapping videos followed [21]. The community was subsequently closed due to the sharing of pornographic content.

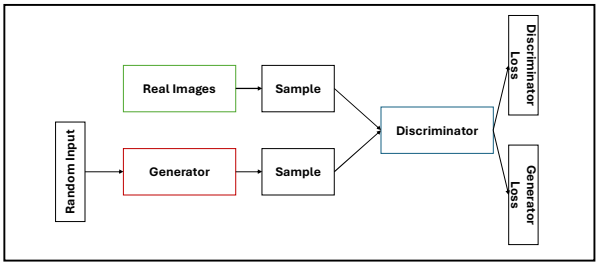

Fig. 1. Illustration of Generative Adversarial Network [17]

The next iteration of deepfakes was created merely for entertainment purposes, however, the results were artificial and characterised by low-quality faces, colour contrast between superimposed faces, visual boundaries, and strange artefacts among sequential frames [22, 23]. The release of several open-source tools, such as FakeApp, DeepFaceLab, and FaceSwap [24], made deepfake creation more accessible to the everyday user. These tools use autoencoders based on the encoder-decoder architecture to create the deepfakes [25]. Autoencoders not only improved the performance and speed of creating deepfakes but also improved the quality.

Further technological advancements have led to improvements in face reenactment, voice cloning, and lip-syncing. Such improvements have led to the creation of extremely realistic deepfakes, which have now initiated the detection arms race. Efforts to detect deepfakes have gained momentum with the release of datasets such as FaceForensics++ [26] and DeepFake Detection Challenge (DFDC) [27]. The FaceForensics++ dataset consists of 1000 original video sequences, manipulated using Deepfakes, Face2Face, FaceSwap and NeuralTextures methods [26]. The DFDC dataset is currently considered the largest, with over 100,000 videos developed using several Deepfake, GAN-based, and non-learned methods and sourcing input from 3,426 paid actors [27]. However, as deepfakes become multimodal, manipulating both visual and audio components, alternative detection techniques must be explored.

B. Temporal Attack Model

With cyberattacks, adversaries perform malicious activities systematically as time progresses [28]. This is also true for cyberattacks involving deepfakes. Therefore, it becomes possible to present cyberattacks incorporating deepfakes using a temporal attack model. The model, introduced by Van Heerden et al. [29], describes the sequence of events and actions an attacker takes over time during a cyberattack. A visual representation of the temporal attack model is presented in Fig. 2.

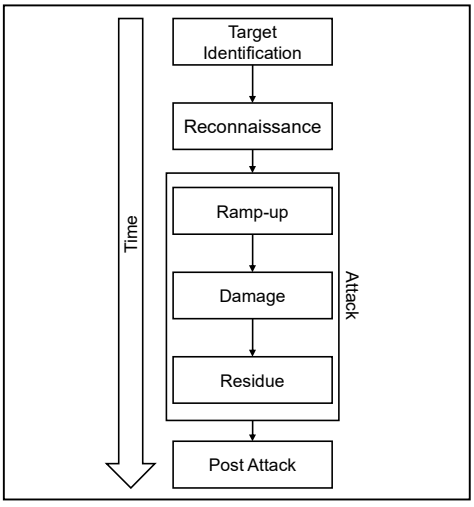

Fig. 2. Temporal Attack Model [27]

The stages of the temporal attack model are [29]:

- Target Identification: represents the selection of the target by the attacker.

- Reconnaissance: involves gathering information about the target by the attacker, specifically exploring for weaknesses or vulnerabilities.
- Attack: using the gathered information and exploiting the identified weaknesses to achieve the end goal.
    - Ramp-up: preparatory actions performed by the attacker.
    - Damage: actions undertaken by an attacker to achieve the final goal.
    - Residue: unintended information remaining after the attack.
- Post Attack Reconnaissance: the attacker verifies the effects of the attack and assesses whether the same methodology can be used again in the future.

The temporal attack model enables a time-oriented analysis of significant deepfake cyberattacks, which is required to develop a comprehensive Deepfake Attack Framework. Mapping these attacks within the temporal model provides a novel perspective that offers deeper insight into the common methodologies employed in deepfake-driven attacks.

III. Timeline of Significant Deepfake Cyberattacks

The increasing number of deepfake cyberattacks in recent years has raised serious concerns regarding the ethical and legal implications of the technology. Several cyberattacks involving deepfakes have successfully demonstrated the misuse of the technology to promote misinformation or assist with identity theft. Creating the necessary awareness to enable improved detection and protection against cyberattacks driven by deepfakes requires a better understanding of past attacks. This section explores the most significant deepfake cyberattacks (see Fig. 3) that have emerged since the technology first appeared in 2017.

A. 2018

The Belgian political party, Flemish Socialist Party, used deepfake technology to develop a fictional address given by United States (US) President Donald Trump. The untrue address called Belgium to follow America and exit the Paris climate agreement. While this particular deepfake video can't be viewed as a cyberattack, the motivation behind the creation of the video was to influence and focus attention on the climate change debate. Various followers of the party fell victim to the deepfake video, even though the intended purpose was not to trick supporters [30].

B. 2019

The first known instance of a scam involving deepfake voice cloning happened in 2019. A Chief Executive Officer (CEO) of an unnamed company based in the United Kingdom (UK) was targeted to transfer €220,000 to a Hungarian supplier. The CEO believed he was communicating with the chief executive of the company's German parent company due to the subtle German accent and expected voice tone (melody). It was only when a follow-up payment was requested that the CEO grew suspicious and detected the attack [31].

C. 2020

An investigation found that a bank in the United Arab Emirates (UAE) was defrauded of $35 million due to a deepfake cyberattack. The branch manager of a UAE bank received a phone call from a familiar-sounding voice, the apparent director of the victim company, requesting the

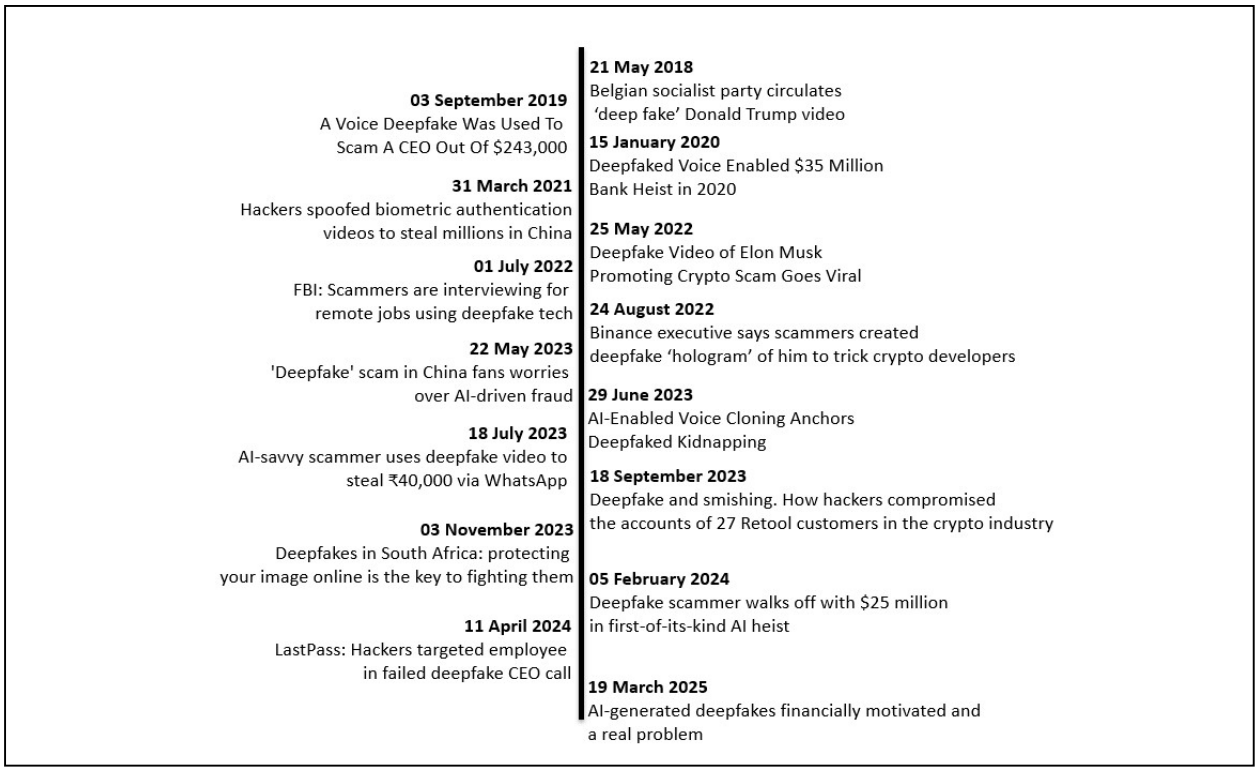

Fig. 3. Timeline of the most significant deepfake cyberattacks

transfer of the funds. Accompanying emails from a lawyer named Martin Zelner convinced the branch manager to proceed with the transaction. It was later confirmed during the investigation that deepfake voice cloning technology was used to imitate the company director's voice [32].

D. 2021

In a slightly different use of deepfake technology, a Chinese government biometric authentication system that uses facial recognition was attacked. The attackers acquired personal data illegally and developed high-resolution images of the targeted individuals mimicking movement (e.g., nodding, shaking, blinking, and opening of their mouths). The enhanced images successfully fooled the biometric authentication system, allowing the attackers to acquire biometric data used to generate fraudulent tax invoices [33].

E. 2022

The year 2022 saw a surge in cyberattacks using deepfake technology. These attacks not only illustrated the improvements in deepfakes but also confirmed the diverse applicability of deepfakes as used in different types of cyberattacks.

One of the first cryptocurrency scams using deepfake technology targeted Elon Musk. While not convincing, the video used a dishonest representation of Elon Musk to promote a cryptocurrency investment scheme with huge returns. Elon Musk was quick to debunk the scam on the Twitter (now called X) social media platform [34].

In July 2022, the Federal Bureau of Investigation (FBI) issued a public warning about scammers using deepfake technology to pose as job applicants during interviews for remote roles. Information Technology (IT) organisations are specifically targeted, with the scammers attempting to acquire positions that provide access to customer personal identifying information (PII), financial data, corporate IT databases and/or proprietary information. Obtaining such information could enable scammers to launch identity fraud schemes. However, the FBI highlighted several flaws still present in the deepfake technology, such as the misalignment of auditory actions with what is visually presented [35].

Another cryptocurrency-related cyberattack took place in 2022, with attackers creating a deepfake of Patrick Hillmann, the chief communications officer at Binance at the time of the attack. The deepfake, described as an "AI hologram", was created to trick individuals into a meeting with a Binance official to discuss opportunities of listing tokens on the Binance cryptocurrency platform. The scam only became known once a targeted individual asked Patrick Hillmann to confirm the meeting took place [36].

F. 2023

Deepfake-related cyberattacks continued in 2023, with the first known cyberattack targeting South African citizens emerging towards the end of 2023.

In May 2023, news agencies reported on a sophisticated cyberattack using deepfake technology that targeted a man in the city of Baotou, in the region of Inner Mongolia. The attacker, posing as a friend, used deepfake technology to launch AI-driven fraud against the target. Using manipulated voice and facial data, the attacker convinced the target to transfer $622,000. The target only became aware of the fraudulent activity when the friend was unable to confirm receipt of the transferred funds [37].

In a different extortion attempt, AI-enabled voice cloning was used as part of a virtual kidnapping incident. An attacker attempted to extort $1 million from an Arizona-based woman, claiming he had abducted her daughter. The mother was convinced due to the frantic pleas and cries heard in the background, which sounded identical to her daughter's voice, while she engaged with the attacker over the phone. Local police discovered that the alleged kidnapping was a scam and confirmed the use of deepfake voice-cloning technology [38].

Similar to the cyberattack in May 2023, a retired government employee from India fell victim to a scam driven by deepfake technology. The target received a call from a person pretending to be his friend and a past colleague. Camaraderie was established and confirmed via WhatsApp conversations. Another call followed from the attacker, requesting an advance of 40,000 Indian Rupees for his sister-in-law's emergency surgery. Initially doubtful, the target did not immediately proceed to transfer the funds but was convinced following a video call from the attacker. It was only when an additional advance of a further 35,000 Indian Rupees was requested that the target decided to first confirm with his friend, who was unaware of the request and failed to confirm receipt of the early transferred funds [39].

Retool, a software development company, was targeted by an SMS-based social engineering attack that compromised 27 cloud customer accounts. The victims were sent personalised SMS messages from an attacker impersonating an IT department representative, claiming to address a payroll issue. The message contained a link directing them to a counterfeit login page featuring multi-factor authentication (MFA). To acquire the MFA codes, the attacker used an AI-generated deepfake of a familiar voice from the IT team. Once the MFA codes were obtained, the attacker gained access to the victims' GSuite accounts [40].

2023 concluded with one of the first known deepfake-related cyberattacks targeted at South Africans. Leanne Manas, a well-known South African businesswoman and news presenter, fell victim to a deepfake-related cyberattack. The developed deepfake, which appeared as part of false news stories and fake advertisements, used manipulated videos of Leanne Manas to promote products and get-rich-quick schemes [41].

G. 2024

While a few deepfake-related cyberattacks were publicly reported in 2024, the sophistication of these attacks had improved.

According to a report by the South China Morning Post, a Hong Kong-based multinational corporation suffered a substantial financial loss as a result of a deepfake-driven scam. The sophisticated cyberattack involved a video conference in which every participant, except the targeted employee, was a digitally manipulated replica of actual staff members. Among these fabricated figures was a deepfake version of the company's chief financial officer, who instructed the victim to transfer $25.6 million. The fraud remained undiscovered until a week later. This incident marks the first reported cyberattack featuring multiple deepfake individuals [42].

An employee of LassPass was reportedly targeted by an attacker using deepfake audio to impersonate LastPass's chief

executive officer. The targeted employee became suspicious due to the use of an uncommon business channel (WhatsApp) and communication outside standard business hours. While the cyberattack was unsuccessful, the attack confirmed an increasing use of AI-generated deepfakes to impersonate executives in fraud campaigns [43].

H. 2025

The South African Parliament's YouTube accounts were compromised to promote a fraudulent cryptocurrency token called "$Ramaphosa". Unauthorised posts featuring a graphic advertising the presale of the new cryptocurrency token were uploaded to the YouTube channel. This attack shows the growing misuse of well-known individuals to abuse public trust [44].

IV. Deepfake Attack Framework

The presented cyberattacks illustrate the growing sophistication and misuse of deepfake technology. Improving or building defences against deepfake cyberattacks requires understanding the typical methodology, which can be achieved by formulating a novel attack framework. An attack framework typically provides a systematic approach for describing and analysing the progression of cyberattacks. This formal structure can assist in the detection and prevention of such attacks. The cyberattacks posed by deepfake technology, as presented in Section III, highlight the need for innovative strategies to equip government, academia, and the business sector personnel with the fundamental knowledge to detect and defend against such attacks.

This section presents the Deepfake Attack Framework (see Fig. 4), constructed using the key trends and characteristics identified from the evaluated cyberattacks. The key trends and characteristics offer the needed insight to separate and map the different stages of deepfake cyberattacks according to the temporal model, of which the outcome is the attack framework. The framework is tailored to assist professionals in improving their defences against deepfake-based cyberattacks and to create awareness regarding the methodology often followed by such attacks.

A. Stage 1: Target Identification

Initiation of any deepfake cyberattack involves the identification of a target, which can be one or more individuals (representing a single organisation). The choice of target is determined by the underlying motive of the cyberattack. While the evaluated deepfake cyberattacks in Section III revealed financial fraud as the main motivator, other factors, such as disinformation campaigns and causing reputational damage, can also drive deepfake-related cyberattacks. Target(s) are assessed in terms of the ability to impersonate. This is determined by public exposure of the identified target(s) and the availability and/or quality of related voice/video data. Final selection of the target(s) directly depends on the exploitability of visual attributes (voice, appearance and behaviour).

Sudden or unusual attention toward a specific individual or organisation, and an increase in phishing (especially spear phishing) campaigns, can be an early indicator of preparatory actions for a deepfake cyberattack.

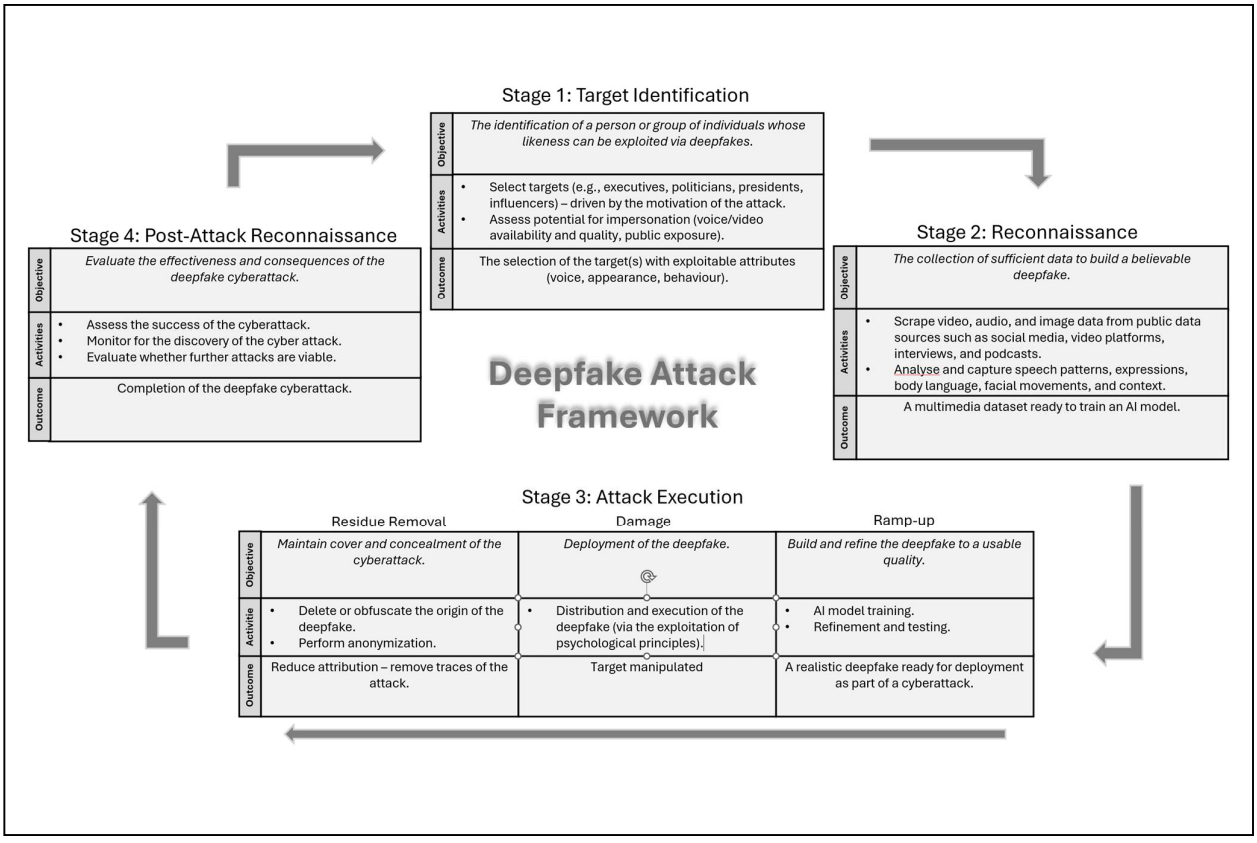

Fig. 4. Deepfake Attack Framework

B. Stage 2: Reconnaissance

Following the target identification is the reconnaissance stage, which involves collecting sufficient multimedia data to develop and build believable deepfake content. Gathering the required data relies on open-source intelligence (OSINT) techniques, deployment of web scrapers, and media downloaders to acquire the necessary video, audio and image data from public sources. In addition, supplementary tools, such as AI voice cloning and facial synthesis, can be deployed to extend the existing dataset. The outcome of the reconnaissance stage is the development of a multimedia dataset to be used to train the AI model.

C. Stage 3: Attack Execution

The attack execution stage follows three sequential steps: ramp-up, damage, and residue removal. The ramp-up step focuses on developing the AI model to produce the required deepfake. Either open-source (DeepFaceLab or FakeApp) or commercial tools can be deployed to assist with the development of the deepfake. Multiple rounds of testing the realism and believability of the developed deepfake will be required to ensure proper refinement. The ramp-up step concludes once a realistic deepfake has been created.

The damage step uses the developed deepfake to conduct the attack. Firstly, the deepfake must be distributed to the intended target. Distribution of the deepfake can occur via various channels, but the most appropriate channel will directly depend on the identified target. The evaluated deepfake cyberattacks revealed common channels to distribute deepfakes, including phone calls, messages (e.g., WhatsApp), video conferencing, and email. Several attacks have shown that the distribution may also involve the real-time generation of the deepfake, especially during phone calls or virtual meetings. Exploiting psychological principles to manipulate individuals, such as the misuse of authority, urgency, familiarity, or reciprocity, is often also displayed during the attack. The purpose of the deepfake is to execute the malicious objective of the attacker. Most commonly, as disclosed by the review of the known deepfake cyberattacks, attackers are driven by financial greed. The result, if the deepfake is successful, is the manipulation of the target, possibly resulting in financial loss, reputation damage, or public deception.

The final step of the attack execution stage is residue removal, referring to unintended information remaining after the damage step has been completed. To ensure the success of the attack and avoid discovery, further attempts to conceal the attack are often performed. Previous deepfake cyberattacks that were financially motivated showed attackers attempting to conceal the fraudulent activities by transferring stolen funds between multiple points. In addition, attackers will also endeavour to cover or remove traces of the attack. This often involves deleting or obfuscating the origin of the deepfake. Using anonymising tools or platforms (e.g., creating and using fake personas) further assists in obscuring the attack. Trace removal reduces attribution and increases the success of the attack.

D. Stage 4: Post-Attack Reconnaissance

The post-attack reconnaissance stage follows the conclusion of the attack. The main focus of this stage is to evaluate the success of the attack (e.g., was the developed deepfake cyberattack believable). Depending on the attack, as formed by the motivation of the attacker, various metrics are evaluated to determine the success of the attack. For example, with financially motivated cyberattacks, success is confirmed with the transfer of the requested funds. For cyberattacks driving misinformation or disinformation campaigns, impact is gauged through target response and behaviour (e.g., did the target(s) fall for the deepfake). The attacker will also monitor available sources for the possible discovery of the cyberattack. An undiscovered deepfake cyberattack confirms the success and completion of the attack. As a final step, the attackers will evaluate if additional attacks are viable (e.g., further exploiting the exposed confusion). This often causes the exposure and discovery of the original attack (e.g., this is often witnessed by cyberattacks targeting individuals), which is an unintended side effect of the post-attack reconnaissance stage.

V. Discussion

The previous section introduced the Deepfake Attack Framework. The purpose of the Deepfake Attack Framework is to offer valuable insights required to understand the risks associated with deepfakes. The Deepfake Attack Framework presents a theoretical solution that can be applied practically to minimise risk and enable personnel to be better prepared to defend against deepfake-driven cyberattacks. This section further discusses the developed framework in terms of its effectiveness as a detection strategy, as well as the consequences of deepfake cyberattacks.

A. Effectiveness of Detection Strategies

As noted in Section I, most deepfake detection research has centred on the application of AI techniques to differentiate between authentic and manipulated content. However, these solutions are not easily accessible or widely available to the general public. As a result, individuals must depend on their own awareness and critical thinking skills to identify cyberattacks involving deepfakes. Two studies [45, 46] have found that humans can outperform certain AI models in deepfake detection. This is due to humans' improved ability to visually process faces, which has created a strong belief in the "wisdom of the crowd" regarding deepfake detection [45, 46]. However, prior research has shown that humans tend to be biased and more often classify deepfakes as authentic [15].

The developed Deepfake Attack Framework can assist in overcoming such bias by stressing the risks associated with deepfake cyberattacks. The Target Identification and Reconnaissance stages highlight the danger of overexposing individuals, especially with the proliferation of social media. Awareness campaigns should emphasise such risks and encourage minimal exposure of personal content that could be used to develop deepfakes. While an online presence is often required for professionals, organisations should consider techniques, such as watermarking, to discourage the misuse of multimedia showcasing their employees. The Attack Execution stage, and more specifically, the damage step, often relies heavily on exploiting psychological principles. While the exploitation of psychological principles is common practice with traditional social engineering attacks, e.g., phishing, both the established framework and past attacks evaluated confirmed that such practices are also true for deepfake cyberattacks. The Deepfake Attack Framework highlights the importance of increased vigilance when encountering visual or auditory requests that exploit common psychological tactics designed to manipulate individuals. Verification of requests across multiple channels must become standard practice. Although the Post-Attack

Reconnaissance stage indicates the completion of a deepfake cyberattack, the Deepfake Attack Framework show further attacks remain a possibility. Individuals must remain attentive, especially to repetitive requests, and corroborate such requests via appropriate channels.

It is essential for existing awareness campaigns to incorporate the established Deepfake Attack Framework to enhance awareness that humans must not accept all things at face value and that living in the 21st century requires deeper evaluation of visual or auditory information.

B. Deepfake Consequences

It is important to note that the consequences of deepfake cyberattacks stretch farther than just the intended attack. From the target's perspective, such an attack can have interpersonal consequences. While few studies have been conducted, research has shown that deepfakes can modify an individual's memories or even implant false memories [47]. Furthermore, deepfakes can also alter one's attitude towards the target of a deepfake [47]. Another important consequence not yet widely explored is the impact on the nonconsensual victim, the individual portrayed in the deepfake. In a deepfake, the voice and/or visual presentation of the victim is altered to do or say something not done or said before. Such severe alterations can have lingering effects of mistrust, social confusion and reputational damage.

The introduction of the Deepfake Attack Framework can assist in raising awareness of the consequences of deepfake cyberattacks. The framework describes the expected attack methodology, and although the impact of such attacks is expected at the Attack Execution stage, the framework shows that implications can occur across the various stages of a deepfake cyberattack. For example, it might not be possible for the cyberattack to proceed past stage 2 (Reconnaissance) due to the inability to build the required dataset, however, the information gathered during stage 1 (Target Identification) could still lead to other social engineering attacks, such as phishing or identity theft. It is possible to deduce from the framework additional consequences that could occur as a direct result of a deepfake cyberattack, such as the reveal of sensitive information, causing breaches or insider attacks. If the victim or target, either an individual or an organisation, fails to prevent a deepfake-based cyberattack, they could face lawsuits, fines, or damage to their public standing.

VI. Conclusion

Deepfakes have garnered widespread attention due to the use of the technology in cyberattacks to influence, spread disinformation, or perform fraudulent activities. This paper evaluated the most significant deepfake cyberattacks with the intended goal of developing a Deepfake Attack Framework. The framework captures the expected methodology of a cyberattack driven by deepfakes. The framework captures the attack methodology across four distinct stages: Target Identification, Reconnaissance, Attack Execution, and Post-Attack Reconnaissance. While the Deepfake Attack Framework presents a theoretical solution, the framework can be applied practically to minimise risk and enable personnel from government, academia, and the business sector to be better prepared to detect and defend against deepfake-driven cyberattacks. It is, therefore, important for this framework to be incorporated into cybersecurity awareness campaigns and training sessions. Regular exposure will improve awareness and reduce the success of deepfake cyberattacks. Future work will focus on maturing and expanding the framework as additional deepfake cyberattacks are evaluated.

References

[1] S. Patil and H. Shankar, "Transforming healthcare: harnessing the power of AI in the modern era," in International Journal of Multidisciplinary Sciences and Arts, vol. 2, no. 2, pp.60-70, 2023.

[2] C. M. Ibegbulam, J. A. Olowonubi, S. A. Fatounde, and O. A. Oyegunwa, "Artificial intelligence in the era of 4IR: drivers, challenges and opportunities," in Engineering Science & Technology Journal, vol. 4, no. 6, pp. 473-488, 2023.

[3] S. J. Plathottam, A. Rzonca, R. Lakhnori, and C. O Iloeje, "A review of artificial intelligence applications in manufacturing operations" in Journal of Advanced Manufacturing and Processing, vol. 5, no. 3, pp. 1-19, 2023.

[4] M. Roshanaei, M. R. Khan, and N. N. Sylvester, "Enhancing cybersecurity through AI and ML: Strategies, challenges, and future directions," in Journal of Information Security, vol. 15, no. 3, pp. 320-339, 2024.

[5] N. Veerasamy and H. Pieterse, "Rising above misinformation and deepfakes," in International Conference on Cyber Warfare and Security, 2022, pp. 340-348.

[6] R. U. Maheshwari, B. Paulchamy, V. Selvaraj, and N. N. Saranya, "Deepfake detection using integrate-backward-integrate logic optimization algorithm with CNN," in International Journal of Electrical and Electronics Research, vol. 12, no. 2, pp. 696-710, 2024.

[7] S. Alanazi, S. Asif, A. Caird-daley, and I. Moulitsas, "Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses" in Human-Intelligent Systems Integration, 2025, pp. 1-23.

[8] S. H. Al-Khazraji, H. H. Saleh, A. I. Khalid, and I. A. Mishkhal, "Impact of deepfake technology on social media: Detection, misinformation and societal implications," in The Eurasia Proceedings of Science Technology Engineering and Mathematics, 2023, pp. 429-441.

[9] "Sumsub Identity Fraud Report 2023," Sum and Substance Ltd, UK, 2023. Available at: https://sumsub.com/files/sumsub_identity_fraud_report_2023.pdf (20 April 2025).

[10] "South Africa shows a 1200% increase in deepfake fraud," Technews Publishing. Available at: https://www.securitysa.com/21083r (20 April 2025).

[11] "Sumsub Identity Fraud Report 2024," Sum and Substance Ltd, UK, 2023. Available at: https://sumsub.com/files/sumsub_identity_fraud_report_2024.pdf (10 June 2025).

[12] B. Neethling, "The technology that presents a significant threat to South Africa," MyBroadband. Available at: https://mybroadband.co.za/news/security/590019-the-technology-that-presents-a-significant-threat-to-south-africa.html (20 April 2025).

[13] L. A. Passos, D. Jodas, K. A. Costa, L. A. Souza Júnior, D. Rodrigues, J. Del Ser, and J. P. Papa, "A review of deep learning ‐ based approaches for deepfake content detection," in Expert Systems, vol. 41, no. 8, pp.1-84, 2024.

[14] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," in Information Fusion, vol. 64, pp. 131-148, 2020.

[15] K. Somoray, and D. J. Miller, "Providing detection strategies to improve human detection of deepfakes: An experimental study," in Computers in Human Behavior, vol. 149, pp. 1-8, 2023.

[16] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), 2018, pp. 1-6.

[17] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in SIGGRAPH97: The 24th International Conference on Computer Graphics and Interactive Techniques, 1997, pp. 353-360.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, vol. 27, pp. 1-9, 2014.

[19] H. Raj, “The Astonishing World of Generative Adversarial Networks (GANs),” Medium. Available at: https://medium.com/@henilsinhrajraj/the-astonishing-world-of-generative-adversarial-networks-gans-92e29423dc13 (21 April 2025).

[20] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F. Y. Wang, “Generative adversarial networks: introduction and outlook,” in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 588-598, 2017.

[21] T. D. Fikse, “Imagining deceptive Deepfakes: an ethnographic exploration of fake videos,” Master's thesis, 2018.

[22] A. Shaji George, and A. S. Hovan George, “Deepfakes: The Evolution of Hyper realistic Media Manipulation,” in Partners Universal Innovative Research Publication, vol. 1, no. 2, pp. 58–74, 2023.

[23] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, “Deepfakes evolution: Analysis of facial regions and fake detection performance,” in International Conference on Pattern Recognition, 2021, pp. 442-456.

[24] J. Botha and H. Pieterse, “Fake news and deepfakes: A dangerous threat for 21st century information security,” in ICCWS 2020 15th International Conference on Cyber Warfare and Security, 2020, pp. 57-67.

[25] T. Fernando, D. Priyasad, S. Sridharan, A. Ross, and C. Fookes, "Face Deepfakes--A Comprehensive Review," To be published, Available at: https://arxiv.org/pdf/2004.07532 (22 April 2025).

[26] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1-11.

[27] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, “The DeepFake detection challenge dataset,” Arxiv. Available at: https://arxiv.org/pdf/2006.07397 (22 April 2025).

[28] M. R. Rahman, B. Wroblewski, Q. Matthews, B. Morgan, T. Menzies, and L. Williams, “Mining Temporal Attack Patterns from Cyberthreat Intelligence Reports,” Arxiv, Available at: https://arxiv.org/pdf/2401.01883 (22 April 2025).

[29] R. Van Heerden, H. Pieterse, and B. Irwin, “Mapping the most significant computer hacking events to a temporal computer attack model,” in ICT Critical Infrastructures and Society: 10th IFIP TC 9 International Conference on Human Choice and Computers, 2012, pp. 226-236.

[30] H. Von der Burchard, “Belgian socialist party circulates ‘deep fake’ Donald Trump video,” Politico. Available at: https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/ (24 April 2025).

[31] J. Damiani, “A Voice Deepfake Was Used To Scam A CEO Out Of $243,000,” Forbes. Available at: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=17b2110c2241 (24 April 2025).

[32] M. Anderson, “Deepfaked Voice Enabled $35 Million Bank Heist in 2020,” Unite.AI. Available at: https://www.unite.ai/deepfaked-voice-enabled-35-million-bank-heist-in-2020/ (24 April 2025).

[33] J. Nash, “Hackers spoofed biometric authentication videos to steal millions in China,” BiometricUpdate.com. Available at: https://www.biometricupdate.com/202103/hackers-spoofed-biometric-authentication-videos-to-steal-millions-in-china (24 April 2025).

[34] M. Di Salvo, “Deepfake Video of Elon Musk Promoting Crypto Scam Goes Viral,” Decrypt Media. Available at: https://decrypt.co/101365/deepfake-video-elon-musk-crypto-scam-goes-viral (24 April 2025).

[35] M. Kan, “FBI: Scammers are interviewing for remote jobs using deepfake tech,” Mashable, Inc. Available at: https://mashable.com/article/deepfake-job-interviews-fbi (24 April 2025).

[36] L. Hurst, “Binance executive says scammers created deepfake ‘hologram’ of him to trick crypto developers,” Euronews. Available at: https://www.euronews.com/next/2022/08/24/binance-executive-says-scammers-created-deepfake-hologram-of-him-to-trick-crypto-developer (24 April 2025).

[37] “‘Deepfake' scam in China fans worries over AI-driven fraud,” Reuters. Available at: https://www.reuters.com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/ (24 April 2025).

[38] J. Vijayan, “AI-Enabled Voice Cloning Anchors Deepfaked Kidnapping,” Dark Reading. Available at: https://www.darkreading.com/cyberattacks-data-breaches/ai-enabled-voice-cloning-deepfaked-kidnapping (24 April 2025).

[39] S. Jain, “AI-savvy scammer uses deepfake video to steal ₹40,000 via WhatsApp,” Business Insider. Available at: https://www.businessinsider.in/tech/news/ai-savvy-scammer-uses-deepfake-video-to-steal-40000-via-whatsapp/articleshow/101849089.cms (24 April 2025).

[40] P. Paganini, “Deepfake and smishing. How hackers compromised the accounts of 27 Retool customers in the crypto industry,” Securityaffairs. Available at: https://securityaffairs.com/150981/hacking/retool-smishing-attack.html (24 April 2025).

[41] L. Van Gensen, “Deepfakes in South Africa: protecting your image online is the key to fighting them,” Daily Maverick. Available at: https://www.dailymaverick.co.za/article/2024-03-04-deepfakes-in-south-africa-protecting-your-image-online-is-the-key-to-fighting-them/ (24 April 2025).

[42] B. Edwards, “Deepfake scammer walks off with $25 million in first-of-its-kind AI heist,” Ars Technica. Available at: https://arstechnica.com/information-technology/2024/02/deepfake-scammer-walks-off-with-25-million-in-first-of-its-kind-ai-heist/ (24 April 2025).

[43] S. Gatlan, “LastPass: Hackers targeted employee in failed deepfake CEO call,” Bleeping Computer. Available at: https://www.bleepingcomputer.com/news/security/lastpass-hackers-targeted-employee-in-failed-deepfake-ceo-call/ (25 April 2025).

[44] C. Treger, “AI-generated deepfakes financially motivated and a real problem,” ITweb. Available at: https://www.itweb.co.za/article/ai-generated-deepfakes-financially-motivated-and-a-real-problem/VgZeyvJlwR2MdjX9 (25 April 2025).

[45] M. Groh, Z. Epstein, R. Picard, and C. Firestone, “Human detection of deepfakes: A role for holistic face processing,” in Journal of Vision, vol. 21, no. 9, pp. 2390-2390, 2021.

[46] M. Groh, Z. Epstein, C. Firestone, and R. Picard, “Deepfake detection by human crowds, machines, and machine-informed crowds,” in Proceedings of the National Academy of Sciences, vol. 119, no. 1, 2022.

[47] J. T. Hancock and J. N. Bailenson, “The social impact of deepfakes,” in Cyberpsychology, behavior, and social networking, vol. 24, no. 3, pp. 149-152, 2021.