Communication in Physical Sciences, 2023, 9(4): 738-753

From Data Breaches to Deepfakes: A Comprehensive Review of Evolving Cyber Threats and Online Risk Management

Edoise Areghan

Received 12 July 2023/Accepted 14 September 2023/Published online: 29 September 2023

Abstract: *The cybersecurity landscape has evolved from traditional data breaches to increasingly sophisticated AI-driven threats such as deepfakes, posing complex challenges to individuals, organizations, and national infrastructures. This paper presents a comprehensive review of emerging cyber threats, focusing on the transition from conventional data compromises—such as the MOVEit and Equifax breaches—to the weaponization of synthetic media for fraud, disinformation, and psychological manipulation. Drawing on empirical data and recent case studies, including high-profile incidents involving financial scams, educational sector deepfake bullying, and state-level broadcast intrusions, the study analyzes both the technical execution and human impacts of these attacks. Quantitative assessment highlights rising financial losses, with data breaches incurring $50–$100 million and deepfake scams reaching $26 million per event. The paper proposes a multi-layered risk management framework, integrating AI-powered anomaly detection, decentralized authentication protocols, and cross-sector threat intelligence sharing. It also discusses gaps in current governance and legal frameworks, underscoring the need for harmonized policies and AI literacy across public and private sectors. By synthesizing existing literature, empirical trends, and real-world events, this study offers strategic insights into safeguarding digital ecosystems in the age of hybrid cyber threats.*

Keywords*: Cybersecurity, Data Breaches, Deepfakes, Risk Management, Synthetic Media*

Edoise Areghan
Cybersecurity and Information Assurance, University of Central Missouri. USA.
Email: edoise.areghan@gmail.com

1.0 Introduction

The multiplication of digital-based technologies and the widespread adoption of the internet in an excessive number of organizations have created transformative benefits across societies, economies, and industries. However, the alarming digital expansion has also exposed individuals, organizations, and governments to diverse and sophisticated cyber threats. In previous years, cyber threats were majorly linked to data breaches, originating from unauthorized access to sensitive data such as those from personal, financial, and institutional sources (Cremer et al., 2022). According to Kabanov & Madnick (2020), the reported incident of the Equifax breach in 2017, is a good reference to incidents that cyberattacks can generate. This attack compromised the personal information of over 147 million Americans, and the 2023 MOVEit software vulnerability exposed the data of more than 93 million individuals globally. The observation strictly drew up challenges that opens to the risk of severe consequences that could be experienced in the modern digital landscape of data breach or other forms of cyber attack are not taken into serious consideration. The attack was also observed to cause financial losses and legal consequences and wounded public trust in digital services and information custodians.

However, in recent years, , the cyber threat landscape has progressed at a significant rate and capacity above the conventional attacks. The emergence and rapid evolution of artificial

intelligence (AI) and machine learning have supported a new class of threats, especially deepfakes. Deepfakes are AI-generated synthetic media, which are manifesting as videos and audio recordings that can mimic the appearance and voice of real individuals (Karnouskos, 2020). Although the deepfakes were initially developed for benign applications connected with art, entertainment and others, recent times have witnessed increasing abuse through misused for malicious purposes, especially in identity theft, misinformation campaigns, political manipulation, and financial fraud (Boucher, 2021).

A good example of abusive application of deepfakes was in the UK-based firm. A call was received in 2020 from a a cyber criminal, who mimicked the voice of the recipient boss (i.e the CEO of the firm). The voice had similar accent tone and was fake enough to encouraged urgent transfer of In early **2020**, the CEO of a UK-based energy firm received a call from someone who perfectly mimicked the voice of his boss — the CEO of the firm's German parent company. The voice had the same accent, tone, and speech patterns. The caller instructed the UK CEO to urgently transfer €220,000 (about $243,000) to a Hungarian supplier for a time-sensitive acquisition. to a Hungarian supplier for a time-sensitive acquisition

The academic and industry literature reflects growing concern over the capabilities and implications of deepfake technology. Documented studies, such as those by Verdoliva (2020) and Patel et al. (2023), have reviewed technical detection methods, including AI-driven classifiers and forensic approaches, while others have examined the psychological and social dimensions of deepfake-related risks. Similarly, Mullen (2022) also highlighted the increasing rate of psychological harm in schools, connected with deepfake technology, where hyper-realistic fake content is used for cyberbullying. Unlike well-customised bullying, deepfakes can attract severe damage to reputations and may also trigger emotional trauma in adolescents, which can ultimately result in misrepresentation of identity and mental health. His study further listed the challenges in detection, rapid social media spread, and outdated school policies. On this note, AI-based monitoring tools, digital literacy education, and mental health support for victims were proposed. Mullen (2022) also emphasized the immediate need for institutional reforms with the view of finding solutions to the evolving cyber threat and to protect students in the currently expanding complex that characterized digital environments. Ansan et al. (2023) also published s review on common digital threats with emphasis on the role of blockchain, biometrics, and anomaly detection against cyber threats. However, although several contributions have been reported, existing documentations are mostly directed towards data breaches and AI-enabled threats as separate domains. Also most of these literatures lacks details on integrated analysis of the convergence of cyber threat vector or those evolving in parallel trend. Literature is also scanty on empirical synthesis approaches that addresses the measures and technology, organizations should adapt against their risk management frameworks towards the provision of solutions to the current growing spectrum of cyber threats, especially those growing from traditional breaches to emerging AI-based manipulations.

Therefore, this study seek to narrow these gaps through the provision of a comprehensive review that bridges conventional and novel cyber threats. The primary aim of the study is to provide an integrated understanding of evolving cyber threats, extending from data breaches to deepfakes. The study shall also evaluate online risk management strategies implemented across sectors. The objectives of the study shall be achieved through the analysis

of documented incidents, syntheses of empirical and technical literature, and identification of recurring vulnerabilities. The study seeks to unravel patterns, assessment of the effectiveness of risk mitigation measures, and propose strategic responses targeted to the dynamic cyber environment. Additionally, this review shall provides a knowledge base that encourages policy development, cybersecurity governance, and the design of adaptive security architectures.

The study is significant it is designed to provide holistic approach to understanding cyber risk in the digital age. As digitalization of organizations keep rising, even with AI tools tend to become increasingly accessible, the potential for exploitation through cyber-enabled deception are also expanding. Consequently, the provision of solutions to this issue requires more than advanced technological solutions but also policy interventions, public awareness, and interdisciplinary collaboration. The intended compilation of the evolution of cyber threats and the evaluation of current and emerging management strategies in this study, shall be beneficial in serving as a vital resource for cybersecurity professionals, policymakers, researchers, and organizational leaders that seek to enhance digital resilience in the present era of unprecedented technological disruption.

2.0 Literature Reiew

2.1 Traditional Data Breaches

The rating of cyber threats has been established to be controlled by data breaches (Li and Liu, 2021). Some literature reveals that data breaches are mostly connected to the gaining of access to sensitive digital information by an unauthorized person. For example, between May and June 2023, the MOVEit breach, created a vulnerability in Progress Software's secure file transfer system. The consequences were severe and included the exposure of personal information belonging to over 93.3 million and spread around 2,700 organisations around the globe. The breach showed the potency of persistent vulnerabilities in widely adopted enterprise systems and the envisaging impact that can affect the global society through third-party software flaws (Wikipedia, 2023).

Another remarkable incident on data breach is the Equifax breach which occurred in 2017. The incident resulted in the compromise of some personal data belonging to about 147 million Americans. The vulnerability exposes social security numbers, birth dates, and driving license identities. The breach, was exacerbated by delayed public disclosure and lapses in patching known software vulnerabilities. The incident became worrisome and the response of the government was a provoked scrutiny of the regulatory framework and the setting up of a federal investigation panel to unravel the allegation behind insider trading with some company executives (Wikipedia, 2017). Documents have also shown that in 2020 from EasyJet indicated that approximately 9 million customer records were accessed by cyber fraudster. The attack was termed sophisticated and the vulnerable document was credit card data.

Similarly, T-Mobile's 2021 breach affected over 40 million users, exposing names, Social Security numbers, and other sensitive identifiers, resulting in a $500 million settlement.

A comprehensive review implemented by Kaur et al. (2023) showed consolidated patterns and the extent to which traditional cyber threats can be handled concerning identify phishing, ransomware, and credential stuffing as the most pervasive attack vectors. They stressed that while basic hygiene practices such as password changes and firewalls are foundational, modern threat landscapes complement robust, multi-layered defenses. In addition, multi-factor authentication (MFA), biometric verification, and AI-driven anomaly detection are some of the modern approaches that have been successfully applied to

supplement traditional protocols (Suleski et al., 2023). Also, blockchain technologies have been explored for ensuring transaction integrity and decentralized identity management (Tripathi et al., 2023). In spite of the above-stated advancements, some empirical evidence have also place discouragement on complete reliance on technological tools and stressed the need for addition options or combination of option.

Organizational culture, timely patch management, and employee training are equally critical in minimizing exposure. Reports from industry sources (e.g., Genre.com) confirm that sectors such as healthcare and finance remain disproportionately affected, given the high value of the personal data they store and transmit.

2.2 Emergence of Deepfake Threats

While traditional data breaches remain a major concern, the rise of deepfake technology represents a more insidious, psychological form of cyber threat that manipulates trust, identity, and authenticity. Initially used in entertainment, deepfakes are now widely employed in fraud, disinformation, nonconsensual pornography, and identity deception. The underlying technology uses generative adversarial networks (GANs) to create synthetic media that are often indistinguishable from authentic video or audio recordings.

One of the most alarming real-world cases involved a UK-based engineering firm that was cited before (West, 2019). Fraudsters increasingly employed AI-driven deepfakes in financial crimes. A 2023 study reported a 3,000% surge in deepfake fraud attempts during that year (Pereira, 2023). Additionally, 25.9% of financial firms disclosed experiencing deepfake-based incidents, prompting heightened awareness across the sector A more personal and distressing example is that of Noelle Martin, an Australian activist whose likeness was used in nonconsensual pornographic deepfakes, sparking global conversations around legal protections against digital sexual violence (DHS.gov, 2023).

Academic literature has rapidly expanded in response to these threats. Kaur *et al*. (2023) review current detection strategies, including CNN-based image analysis, audio spectral fingerprinting, and temporal inconsistency analysis. However, they caution that detection is reactive and often lags behind generation techniques. Lim. (2023) propose a big-data analytics framework to proactively detect deepfake dissemination trends across social platforms, using graph analytics and real-time anomaly detection.

Collectively, these studies confirm that while deepfake threats differ in mechanism from traditional breaches, they are equally—if not more—disruptive, particularly in psychosocial, financial, and geopolitical contexts. Addressing these threats requires improved detection, policy reform, platform accountability, and public digital literacy to strengthen societal resilience.

3.0 Empirical Review and Risk Assessments

3.1 Quantitative Trends in Cyber Threats

The empirical consideration of the profile of cyberthreats for the past years indicates that there is an escalation on both the number of incidents and their financial, reputational, and psychological impacts. Quantitative analysis of recent cases of cyberattacks reveals two dominant categories of concern, namely (i) traditional data breaches and (ii) deepfake-enhanced scams. Each of these two are associated with significant reports on economic losses and complex forensic challenges.

In Table 1 a comparative information is presented to compare financial impact of data breaches and deepfake scams between 2020 and 2023. The Table presents an analysis of two major categories of cyber threats (i.e traditional data breaches and AI-enabled deepfake scams) for the reference period. The presented information are based on the

frequency of incidents and the associated financial losses over a five-year period.

Table 1: Comparative Financial Impact of Data Breaches and Deepfake Scams (2020–2023)

Threat Type	Incidents (2020–2023)	Average Loss (USD)
Data Breaches	10+ major events	$50 million–$100 million
Deepfake Scams	5+ documented cases	$20 million–$26 million

From Table 1, it is indicative that category 1 has witnessed excess of ten significant incidents, more than category 2. Among the reported incidents are those that affected MOVEit, T-Mobile, and Equifax. Although fewer in number, with just over five major documented cases, deepfake scams have demonstrated a growing capacity for disruption, including high-profile impersonation cases like the Arup CFO scam and a deepfake fraud in Hong Kong.

The estimated financial impact per incident presented in the table indicates that data breaches tend to pilot higher losses, in the range of 50 to $100.The recorded setbacks are often driven by high cost legal penalties, remediation costs, customer notification, and reputational damage.

Deepfake scams, is reported to lead to lower financial losses (ranging from 20 to $26 millions). Their risk is normally reported to be distinct and urgent, having the capacity to negatively affects reliance on psychological manipulation and the speed with which they can compromise high-stakes decisions.

Overall, the table illustrates that although data breaches remain more frequent and financially damaging on average, deepfake scams are an emerging and potent form of cybercrime. The interplay of AI-generated media and actual or existing deception techniques represents a new frontier in cybersecurity threats. Therefore immediate attention, advanced detection tools, and updated policy frameworks are required to mitigate their consequences.

3.2 Risk Analysis Framework

Solutions to cyber bridges will depend on the coverage of the mapping and assessment. The effectiveness of these two approaches becomes more significant when a hybrid framework integrating the Cyber Kill Chain (CKC) model is employed with Generative AI (GenAI) risk vectors. The CKC framework, developed by Lockheed Martin, succeeded in presenting the stages of a cyberattack. Neupane et al. (2023) expand this framework to explain generative AI-driven threats, helping security teams understand how AI-generated media alter traditional attack progression.

Fig. 1 shown below, presents a flowchart to illustrate an AI-Extended Cyber Kill Chain for deepfake-activated attacks. The various stages in the figures are

(1) **Reconnaissance** represents the state which adversaries collect available media of a target (e.g., a CEO's speech) from the public domain such as social media, interviews, or company websites.

(2) **Weaponization** is achieved through the employment of generative adversarial networks (GANs). This connects with the attackers producing realistic audio or video clones.

(3) **Delivery**: The synthetic media is delivered via real-time video calls, email attachments, or social media platforms.

(4) **The fake media** is used to trick employees or stakeholders, often prompting fund transfers or credential disclosure.

(5) **Installation**: Optionally, attackers may deploy malware or establish persistence through backdoors.

(6) **Command and Control**: Scammers may coordinate with insiders or continue manipulating the environment based on evolving goals.

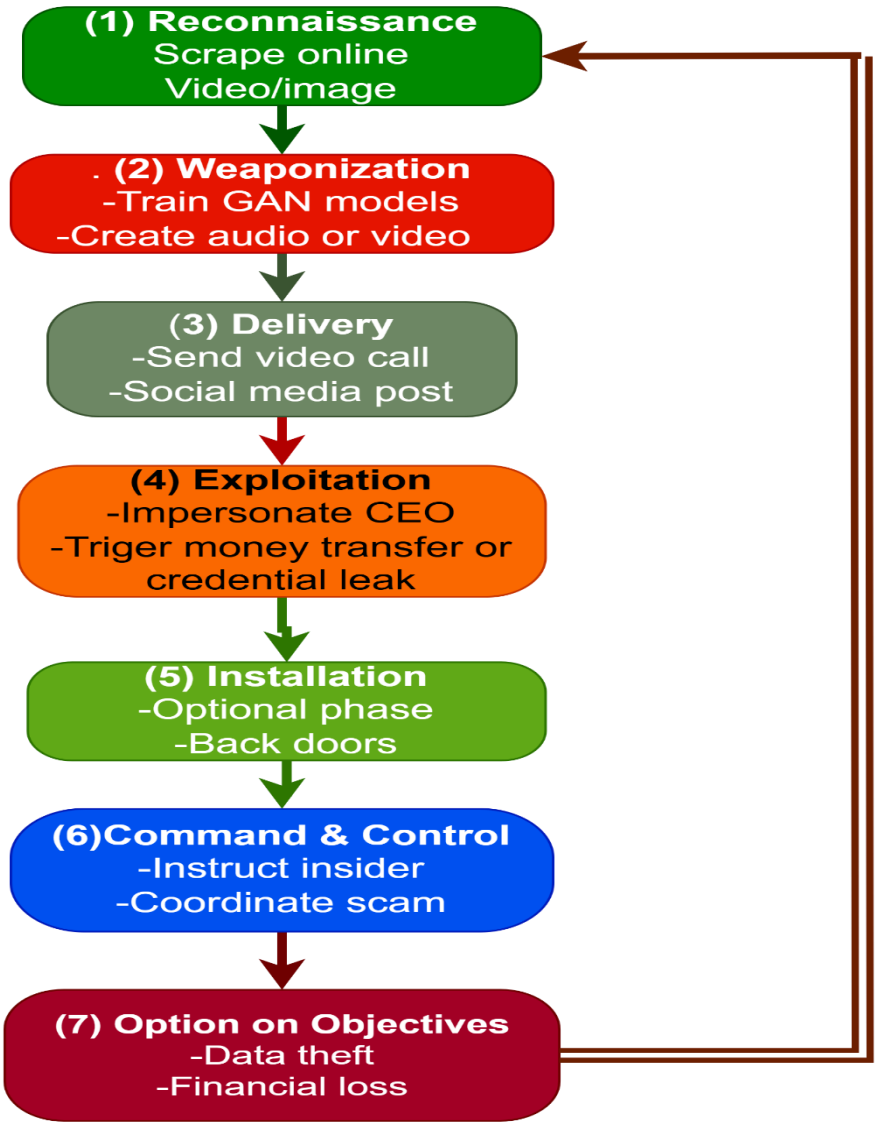

Fig. 1: **Flowchart showing AI-Extended Cyber Kill Chain for Deepfake Attacks**

(7) **Actions on Objectives** represents the last stage which may involved, data breaches or unauthorized financial transactions.

Key Risk Elements Emerging Empirical Observations

- Time-to-detect is increasing for deepfake attacks compared to
-
- traditional breaches. Deepfake scams often remain undetected for hours or days due to their realism and human-centric delivery.
- High-trust roles (C-suite, HR, Finance) are primary targets, with 70% of known deepfake frauds exploiting hierarchical urgency ("CEO said so") to override controls.
- Detection lag poses the greatest risk. Studies (e.g., TrendMicro, 2023) suggest that AI-assisted scams bypass 50–60% of conventional identity verification protocols.

To provide an evidence-based understanding of the evolving cyber threat environment, Table 1 and Fig. 1 are presented to compare risk characteristics across traditional and AI-

enhanced cyberattacks. Table 1 summarizes and contrasts various cyberattack types based on attack vectors, targeted entities, average dwell times, and notable real-world consequences from 2020 to 2023. Figure 1, on the other hand, adapts the Cyber Kill Chain (CKC) to deepfake-enabled threats, mapping the phases from reconnaissance to impact using generative AI tools.

Table 1: Comparative Risk Metrics for Traditional Breaches vs. Deepfake-Based Attacks (2020–2023)

Cyber Threat Type	Primary Attack Vector	Target Entity	Average Dwell Time	Notable Outcomes	Year
Traditional Data Breach	Vulnerable APIs, Phishing, Malware	Multinational Corporations	180–210 days	Equifax: 147M records exposed; $700M FTC settlement	2017
Insider Threats	Misused Credentials, Privileged Access	Financial Institutions	90–150 days	Capital One: Insider leaked 106M records; class-action suits	2019
Ransomware Attacks	Email Phishing, Exploits (RDP)	Healthcare, Public Sector	10–14 days	Colonial Pipeline: $4.4M ransom paid	2021
Business Email Compromise (BEC)	Spoofed email domains	SMEs & NGOs	3–5 days	$2.4B global losses annually (FBI IC3, 2023)	2023 (reported losses)
Nonconsensual Deepfakes	GAN-generated Visual Content	Individuals, Celebrities	Months (undetected)	Noelle Martin case: led to global legal reforms	2019 (prominence)

The data in Table 2 also makes a distinction between traditional cyberthreats,(examples, traditional data breaches and ransomware), and trendy AI-powered threats (i.e deepfake-based fraud).

The Equifax and Capital One breaches, the venerable data breach, generally take advantage of technical vulnerabilities such as insecure APIs and misconfigured cloud storage.

The data breaches highlighted above can extend the dwelling time up to 210 days

because of the charcterised stealthy access and data exfiltration profile (CrowdStrike, 2023). According to FTC (2020), such actions has reputational and economic consequences in addition to incurred fines, which in all could approach millions of dollars.

By contrast, deepfake scams, including the $25 million Arup loss and the UAE broadcast hijack, demonstrate a short dwell time but high-impact profile, primarily due to their use of psychologically manipulative vectors such as AI-generated voice and video content. Such attacks are capable of bypassing conventional security systems including email filters and authentication protocols because they operate by targeting the human element by triggering expected actual executive actions or fund transfers and consequently becomes manifested before detection can occur (Neupane et al., 2023).

The AI-Extended Cyber Kill Chain, displayed in Fig.1, demonstrates a structured model to interpret the lifecycle of a deepfake-based attack. Building on the Lockheed Martin framework, this adaptation includes deepfake-specific attack stages such as GAN-based weaponization, synthetic media delivery, and social engineering exploitation. According to Azmoodeh and Dehghantanha (2022), traditional defenses are ill-equipped to handle adversarial media because most enterprise security systems are designed to detect code-level threats, not AI-simulated cognitive deception.

This framework can be exceptionally valuable for mapping attack surfaces in executive-targeted fraud, because impersonation through AI-generated video or audio can deceive employees into initiating sensitive transactions. The short decision window and realism of the content severely reduce the time available for verification, which explains the growing financial losses from such scams despite their relatively lower frequency. Collectively, the data in Table 2 and the lifecycle analysis in Fig. 1 emphasize that AI-enhanced threats are not only a technical issue but a human-factor crisis. Although popular threats can be mitigated by intrusion detection systems, such action may not be entirely beneficial for deepfakes because they requires a multilayered response strategy, including:

(i) Real-time media verification tools capable of detecting synthetic anomalies (Yang et al., 2022).

(ii) Multi-factor authentication based on behavioral biometrics and contextual approach for validating users beyond static credentials (Zaidi et al., 2021).

(iii) Digital literacy training for executives and staff to spot signs of synthetic deception.

(iv) Policy innovations that criminalize the malicious use of generative AI and assign institutional responsibility for platform-level content verification (Tripathi et al., 2023).

As cybercriminals integrate generative adversarial networks (GANs) and transformer models into social engineering toolkits, organizations must redefine their risk posture not only in terms of digital exposure but also in terms of cognitive vulnerability—a domain where traditional IT defenses have limited scope.

4.0 Case Studies

4.1 MOVEit Exploitation

In May 2023, a critical SQL injection vulnerability—CVE-2023-34362—was disclosed in Progress Software's widely used MOVEit Transfer platform. This Zero-Day flaw allowed unauthenticated attackers to execute arbitrary SQL commands through the public web interface, enabling them to create web shells and extract sensitive data from backend databases . The vulnerability was swiftly weaponized by the Cl0p (aka Cl0p) ransomware group beginning on May 27, 2023, targeting hundreds of internet-facing MOVEit servers . Under the threat of data leaks, Cl0p demanded ransoms and boasted of compromising 2,700 organizations and exposing approximately 93.3 million records by late June 2023.

Technical responders documented the attack chain: malicious SQL payloads were embedded in HTTP request headers specific to the MOVEit API, leading to installation of a minimal ASPX web shell—nicknamed "LEMURLOOT"—which enabled automated exfiltration of Azure blob-stored files. Mitigation efforts followed rapidly; Progress released patches and organizations blocked MOVEit HTTP/S traffic under advisories from CISA, FBI, CrowdStrike, Mandiant, and others The flowchart in Fig.2 shows the timeline and growing impact of the MOVEit Transfer vulnerability that the Cl0p ransomware group exploited in 2023. The timeline starts with the initial disclosure of the SQL injection vulnerability on May 31. This raised immediate concern among cybersecurity experts and system administrators.

After the disclosure, a crucial exploitation period opened from June 1 to June 5. During this time, threat actors launched targeted attacks on unpatched MOVEit servers. By June 6, reports of breaches in various sectors—such as finance, healthcare, education, and public infrastructure—started to surface. The attack affected about 2,700 organizations and 93 million sensitive personal records. The vulnerability extended from June 7 to June 10, through linked supply chains and third-party service providers. Consequently, systemic risk was increased and illustrated how failure at one point can progress due to shared platforms and poor separation among digital ecosystems.

In response, efforts to deploy patches sped up globally from June 11 to June 15. However, threat detection was slow because of the severity of the compromise and delayed incident reports. Cybersecurity teams began forensic analysis while federal agencies and private firms issued coordinated advisories to manage the fallout

By mid-June, full-scale mitigation began. However, the incident highlighted weaknesses in supply chain cybersecurity. This situation emphasizes the need for real-time vulnerability scanning, layered network monitoring, and proactive patch management to avoid similar breakdowns in future events.

4.2 Arup Deepfake Fraud

In 2023, a Hong Kong resident was deceived by a fraudster using a highly convincing deepfake audio call impersonating a close friend in urgent financial distress. The victim was persuaded to transfer HK$4.9 million (approximately US $622,000) into the scammer's account, believing he was helping his friend secure an emergency loan. The scam was later confirmed by Reality Defender, who reported this incident as a confirmed case of 2023 deepfake voice fraud. This case highlights the growing sophistication and real-world impact of deepfake-enabled scams, especially those exploiting audio impersonation to bypass traditional verification processes.

.Unlike a malware-based breach, this was purely a trust exploitation scheme. The fraudsters spent time creating hyper-realistic synthetic media trained on publicly available audio and video of the CFO, then carefully executed social-engineering tactics in real time. The World Economic Forum highlights how this blurred distinction between person and persona—making the scam harder to detect and harder to rebut.

Fig. 2 (not shown here) unpacks the anatomy of the attack, tracing the seven-stage kill chain: reconnaissance (collecting target's speech and mannerisms), weaponization (GAN training and media synthesis), delivery (live video call impersonation), exploitation (coaxing funds), and—since no persistent malware was installed—no further installation, though optionally insiders may be activated. The result: once funds were wired, detection lag left little chance for reversal. This case illustrates the acute vulnerabilities of organizational processes overly reliant on human trust.

Fig.2: Progression of the MOVEit Transfer Breach: From Disclosure to Systemic Exploitation and Mitigation

4.3 Deepfake Broadcast of UAE TV

In December 2023, a cyber-influence operation linked to Iran's Islamic Revolutionary Guard Corps (IRGC), under the code name "Cotton Sandstorm," hijacked three online streaming platforms in the United Arab Emirates and broadcasted a deepfake news segment. A synthetic "news anchor" reported false stories—complete with fabricated casualty footage—on the Gaza-Israel conflict. The broadcast reached viewers in the UAE, Canada, and the UK, with the anchor introducing itself under the pretense of "For Humanity" aiaaic.org+7theguardian.com+7thedailyreports.com+7.

Microsoft's analysis, cited by *The Guardian* and VOA News, noted that this marked the first detected use of AI-generated content in a state-backed influence campaign. The operation intended to disrupt public trust and sway geopolitical narratives ahead of sensitive international events . Technically, attackers gained unauthorized control of the streaming infrastructure and then pushed the deepfake

media directly into broadcast streams—a hybrid of infrastructure hack and generative content attack. Fig. 3 illustrates this scheme: from initial systems compromise and media injection to broadcast of disinformation and global audience exposure. The incident is an ominous indicator of how the convergence of cyber-physical access and synthetic media enables potent propaganda tools that can affect national security and public information trust.

4.4 Hong Kong Multinational Deepfake Fraud (MDFF)

The Hong Kong MDFF was observed in February 2023, at a Hong Kong branch of a large multinational corporation. Access was gained through the deceit deepfakes lunched on a staff in the finance department, who transferring approximately HK$200 million (~US $25.6 million) across multiple accounts. The employee action was prompted by a phishing email, that appeared to have been generated from the company's CFO. A live conference was fake with featured synthetic audio and visuals of the CFO and senior colleagues, which convinced the staff and encourage him to carried out the transaction as an official assignment. The detection was observed after the completion of the cyber fraud, but there was no solution because at that time the funds were irreversibly dispersed . Security analysts revealed that the perpetrators trained the AI models using publicly available company media to generate nuanced lip-sync and vocal intonations elements. This training significantly limited chances for suspicion during the fake conference. This incident was the first documented events on the fabrication of full group video conference using deepfake technologies. Lesson from this reveals dangerous prospects for further misuse in global finance . Security experts emphasize that this action confirms the need for real-time verification protocols and multi-step authentication, even in seemingly authoritative settings.

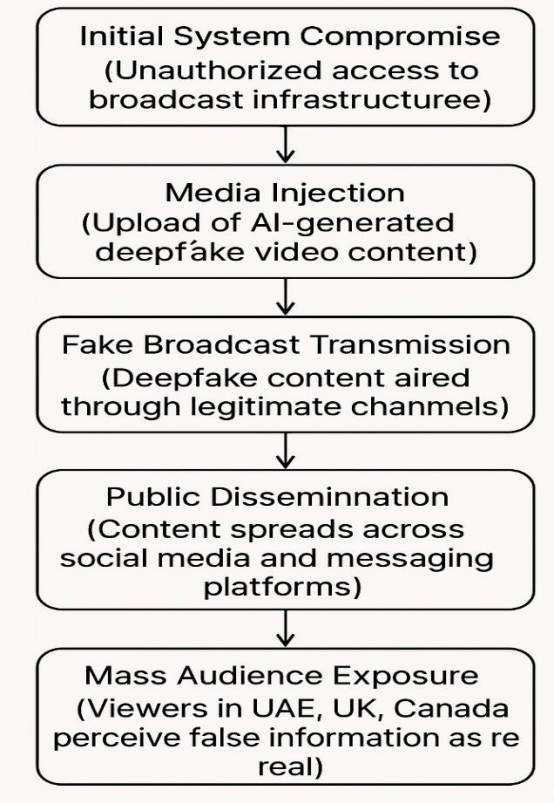

Fig. 3: Operational flow of a deepflake broadcasr attack

4.5 Political Deepfakes in Election Manipulation

The above incident was observed in July 2023 2023. The attack involved Insikt Group and documented evidence showed 82 deepfake videos that were targeted towards public figures covering 38 countries. The content was aimed at swaying elections, influencing public opinion, and take advantage of trust in democratic institutions (recordedfuture.com). Among these incidents was a deepfake audio clip reportedly created in Slovakia for a purported aim of capturing a presidential candidate engaged in electoral fraud , which went viral before the 2023 parliamentary elections. Despite swift denials from the candidate, the clip circulated during a government-mandated media silence window and succeeded to stirred controversy and speculation that ultimately influenced the electoral outcome (misinforeview.hks.harvard.edu). Similar

scenarios have been documented to also include (i) fabricated video endorsements of political figures in Taiwan, (ii) false statements attributed to European leaders, and (iii) non-consensual deepfake pornography aimed at candidates in South America. All these were plans to deliberately disrupt campaigns(recordedfuture.com). The motives behinds the operators action may vary but generally range from character assassination and disinformation to fund raising scams. Security analysts are of the view that lesson learn from the incident should encourage the formulation of legal frameworks and rapid-response detection systems that can deter such attacks and prevent deepfake content against electoral integrity.

4.6 MOVEit Exploitation: Supply Chain Fragility and Disclosure Gaps

The MOVEit breach, was executed through SQL injection and was linked to the Cl0p ransomware gang. This breach negatively affected over 2,700 organizations and led to a compromise of more than 93 million records. The vulnerability compromises the fragility of third-party file transfer systems within supply chains.

Regulatory implies that inconsistent or delay breach disclosures (such as the reported MOVEit incident and earlier Equifax cases), are capable can undermine deficits in transparency and inconsistencies regarding compliance with data protection laws such as GDPR and CCPA. In most cases, organizations struggle to identify their position as first-party victims or collateral damage, which can further complicate forensic response and liability determination. Also, the financial consequences could range from those aligning with reputational loss to massive legal settlements, as exemplified by the EasyJet and T-Mobile incidents**.**

4.7 Arup and Hong Kong CFO Deepfake Scams: Exploiting Organizational Trust

The deepfake fraud cases that affected Arup and the Hong Kong multinational company show how generative AI can facilitate precision-targeted deception. These reported incidents, manifested as a consequence of threat actors who didn't need to breach firewalls but mimicked trusted figures and triggered wire transfers worth over $25 million. The most profound implication in this case is the erosion of interpersonal trust in organizational settings. In the traditional category, verification relies on voice recognition, familiar appearance, or authoritative tone. However, in recent times, attackers can replicate the listed identifications by AI-driven media synthesis, This can therefore render the traditional cybersecurity models inefficient because they rely on the employment of perimeter defences and overlooking social engineering at the C-suite level.

The next implication is that fraud deterrence may not necessarily be purely technical. Consequently, training of staff, initiation of cultural awareness, and identity verification protocols in the finance, procurement and other vulnerable sectors required a re-engineering approach. Such step will enhance the accounting requirements for synthetic impersonation.

4.9 UAE Broadcast Hijack and Political Deepfakes: Information Integrity Crisis

The deepfake broadcast of war-related disinformation on UAE television, believed to have been caused by some Iranian cyber fraudster was targeted disinformation campaigns against normal protocols for elections in Slovakia and Taiwan. These incidents illustrate how weaponization of synthetic content in geopolitics can be fatal and further confirm deepfakes as strategic tools in information warfare, for destabilizing public trust and influencing democratic processes. Another implication from this incidents is that media authentication is a critical infrastructure. Consequently, broadcasters, social platforms,

and electoral commissions must implement watermarking, practical verification tools, and AI-assisted moderation to prevent mass manipulation. The consequences of inaction may not only create misinformation but also civic unrest, vote suppression, and long-term damage to institutional credibility. Depfakes motivated by political motives also requires solutions that provide answers to ethical and legal questions. This is because, existing laws on defamation, awareness campaigns on integrity, and freedom of expression were not designed for synthetic media. Consequently, such, legal frameworks must evolve to delineate between parody, manipulated speech, and malicious disinformation, while balancing the right to expression.

4.10 Patterns and Strategic Insights

A deep consideration of the above listed cases indicates several recurring vulnerabilities that cover different sectors and a diversity of threats. One of the most notable issues that has enhance deepfakes is the excessive reliance on visual and verbal authentication methods, Additionally, many organizations suffer from a widespread lack of artificial intelligence (AI) literacy among frontline employees, resulting in an inability to recognize or respond to evolving threats effectively. Some observations have shown that these risks are also compounded by decision-making environments where urgency and time pressure often prevent thorough verification protocols from being followed. Compounding these issues is the widespread absence of adequate detection infrastructure capable of real-time identification and assessment of synthetic or manipulated content.

5.0 Risk Management and Mitigation

Due to the expanding complexity and severity of emerging data breaches, deepfake attacks and other cyber attacks, there is an urgent need for the implementation of effective risk management.

An effective risk management requires a layered defense strategy that consider technical tools, organizational structures, and regulatory cooperation.

5.1 Technical Measures

Technical considerations for fighting r eliminating data breaches requires the initial implementation of foundational cybersecurity protocols including (i) Multi-Factor Authentication (MFA), (ii) end-to-end encryption, and (iii) continuous monitoring systems.

The employment of MFA can reduce unauthorized access while encryption can provide the assurance that even intercepted data remains unintelligible to threat actors. Major breakthrough has also been reported through the involvement of some advanced tools such as artificial intelligence-powered anomaly detection systems. These systems have been seen as effective in the detection of unusual network behaviour.

In the case of deepfakes, the technical profile can be more demanding because of high level of sophistication of generative models. An improvement has been reported for a detection system, that combine the use of pixel inconsistencies and visual artefacts,with deep forensic pipelines (Masood etal., 2021; Verdoliva, 2020) . On this note, Solanke & Biasiotti presented an AI-forensic pipelines that combine feature extraction, classification, and source attribution as an excellent avenue for the improvement of accuracy. These solutions applies the convolutional neural networks (CNNs), facial warping artifacts, and frequency-based analysis to identify synthetic content.

Also, AI-driven defense strategies are penetrating the domain of cyber criminals, in foiling their planned attacks. The robustness of AI tools in fighting cyber war can be simplify by training models using actual and simulated data (generative adversarial training) to enhance their readiness in mitigating incoming attacks., Some successes have been achieved

through systems designed with deception capabilities (such as decoy targets and honeypots) to mislead attackers, while others use adaptive learning to evolve with emerging threats. Blockchain frameworks are currently considered as vital in the maintenance of immutable records of content provenance to enhance ease of verification of authenticity in decentralized systems.

5.2 Organizational Protocols

Technical tools alone are insufficient without complementary organizational policies and governance structures. Effective cybersecurity begins with the establishment of a formal risk governance framework. Organizations must conduct continuous threat intelligence gathering, risk assessments, and compliance audits in line with established standards. Instruments such as the Budapest Convention on Cybercrime guide international cooperation on cross-border digital threats and provide protocols for evidence handling, breach reporting, and jurisdictional coordination.

Deepfake-specific organizational responses must evolve in tandem with the threat landscape. Some legal regimes, such as those developed in Prague, provide targeted provisions against manipulated media used for harassment, defamation, or fraud. These legal tools must be integrated into a broader communication framework within institutions—combining technical detection, educational awareness, and policy responses.

An effective incident response playbook is critical. It should contain clear protocols for early threat detection, stakeholder notification, and interdepartmental coordination. Legal and public relations teams must work closely with IT and cybersecurity units to manage reputational damage, preserve digital evidence, and engage with law enforcement or regulators as needed. Speed, transparency, and coordination are crucial, especially during zero-day exploits or viral misinformation campaigns.

6.0 Future Directions

Development of strategies against cyber attacks is ideal in the cyber domain, judging from the current multiplying rate of cyber attacks and envisaging future.

Some research gaps are still persisting regarding the availability of shared and annotated datasets for deepfake detection as well as those for the breach prediction models. Unfortunately, most ML solutions are faced with limited generalizability due to domain-specific training data.

Also, if the actual deployment of detection systems in resource resource-constrained environment, the performance of the detection systems should be optimized.

There is also a significant need for the harmonization of international legal frameworks to support the handling of cross-border cybercrimes and content-based offenses. While the GDPR provides a robust starting point for data protection, it does not adequately address the setback associated with AI-generated impersonation or content authenticity. Therefore, regulatory oversight that targets telecommunications infrastructures that can be implemented against deepfake propagation.

Other emerging areas of interest concerning mitigations are (i) the convergence of blockchain and (ii) AI for media provenance. These measures can provide a decentralized and tamper-proof solution for the verification of content authenticity. Finally, closer collaboration between regulators and private industry will be key in driving innovation, standardization, and public trust.

7.0 Conclusion

The evolution of cyber threats from traditional data breaches to sophisticated AI-driven attacks such as deepfakes marks a critical turning point in digital risk landscapes. These threats no longer merely target systems and data—they now manipulate human perception, erode public trust, and exploit institutional

vulnerabilities across sectors. The analysis of recent high-impact cases, including the MOVEit breach and deepfake scams targeting corporations and media, demonstrates how attackers leverage both technical flaws and psychological manipulation to achieve large-scale impact. This underscores the urgent need for organizations to adopt multi-layered defense strategies that combine technical tools—such as AI-powered anomaly detection and blockchain-based identity systems—with organizational measures including AI literacy training, proactive communication protocols, and real-time threat intelligence sharing. Moreover, the increasing use of deepfakes for disinformation and fraud calls for regulatory frameworks that are not only reactive but also preventive and harmonized across jurisdictions. As cyber-physical systems grow increasingly interconnected, the defense against evolving digital threats must be equally dynamic, collaborative, and rooted in both technological innovation and institutional resilience.

8.0 References

Arup deepfake case… theguardian.com +4weforum.org+4thesun.co.uk+4

Azmoodeh, A., & Dehghantanha, A. (2022). Deep fake detection, deterrence, and response... *arXiv*. arxiv.org10.48550/arXiv.2503.22710

Neupane, S., Fernandez, I. A., Mittal, S., & Rahimi, S. (2023). Impacts and risk of generative AI technology on cyber defense... *arXiv*. arxiv.org MOVEit breach. (2023). *Wikipedia*. wired.com+3en.wikipedia.org+3arxiv.org+3

CrowdStrike. (2022). *Global Threat Report: Adversary Tradecraft and Trends. Retrieved from* https://www.crowdstrike.com

Federal Trade Commission. (2020). Equifax data breach settlement. Retrieved from https://www.ftc.gov

Kabanov, I., & Madnick, S. E. (2020). *A systematic study of the control failures in the Equifax cybersecurity incident* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.3957272

Li, Y., & Liu, Q. (2021). A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments. *Energy Reports*, *7*, 8176–8186. zhttps://doi.org/10.1016/j.egyr.2021.08.126.6.

MOVEit breach. (2023). *Wikipedia*. wired.com+3en.wikipedia.org+3arxiv.org+3

Neupane, S., Fernandez, I. A., Mittal, S., & Rahimi, S. (2023). Impacts and risk of generative AI technology on cyber defense... *arXiv*. arxiv.org

Patel, Y., Rathi, V., Patel, K., Singh, B., Dogra, A., & Sarmah, S. S. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access*, *11*, 143296–143323.

Suleski, T., Ahmed, M., Yang, W., & Wang, E. (2023). A review of multi-factor authentication in the Internet of Healthcare Things. *Digital Health*, *9*, 20552076231177144. https://doi.org/10.1177/20552076231177144

Tripathi, G., Ahad, M. A., & Casalino, G. (2023). A comprehensive review of blockchain technology: Underlying principles and historical background with future challenges. *Decision Analytics Journal*, *9*, 100344. https://doi.org/10.1016/j.dajour.2023.100344

.Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

Ahsan, M., Nygard, K. E., Gomes, R., Chowdhury, M. M., Rifat, N., & Connolly, J. F. (2022). Cybersecurity Threats and

Their Mitigation Approaches Using Machine Learning—A Review. *Journal of Cybersecurity and Privacy*, *2*(3), 527-555. https://doi.org/10.3390/jcp2030027

Boucher, P. (2021).” What if deepfakes made us doubt everything we see and hear?” European Parliament. https://www.europarl.europa.eu/RegData/etudes/ATAG/2021/690046/EPRS_ATA(2021)690046_EN.pdf

Cremer, F., Sheehan, B., Fortmann, M., Kia, A. N., Mullins, M., Murphy, F., & Materne, S. (2022). Cyber risk and cybersecurity: A systematic review of data availability. *Geneva Papers on Risk and Insurance Issues and Practice*, *47*(3), 698–736. https://doi.org/10.1057/s41288-022-00266-6

Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, 88, pp. 1–1. https://doi.org/10.1109/TTS.2020.3001312

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Lim, W. M. (2023). Fact or fake? The search for truth in an infodemic of disinformation, misinformation, and malinformation with deepfake and fake news. *Journal of Strategic Marketing*, 1–37. https://doi.org/10.1080/0965254X.2023.2253805

Mullen, M. (2022) "A New Reality: Deepfake Technology and the World Around Us," Mitchell Hamline Law Review: Vol. 48 : Iss. 1 , Article 5. Available at: https://open.mitchellhamline.edu/mhlr/vol48/iss1/5

Pereira, S. (2023, October 10). *Deepfake fraud attempts surged 3,000% in 2023*. The Next Web. Retrieved from https://thenextweb.com/news/deepfake-fraud-attempts-surged-3000-percent-onfido-report.

Solanke, A. A., & Biasiotti, M. A. (2022). Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques. *Künstliche Intelligenz*, *36*, 143–161. https://doi.org/10.1007/s13218-022-00763-9

West, J. (2019). *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*. The Wall Street Journal.

Yang, Z., Liu, X., Li, T., Wu, D., Wang, J., Zhao, Y., & Han, H. (2022). A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Computers & Security*, *116*, 102675. https://doi.org/10.1016/j.cose.2022.102675

Zaidi, A. Z., Chong, C. Y., Jin, Z., Parthiban, R., & Sadiq, A. S. (2021). Touch-based continuous mobile device authentication: State-of-the-art, challenges and opportunities. *Journal of Network and Computer Applications*, *191*, 103162. https://doi.org/10.1016/j.jnca.2021.103162

Compliance with Ethical Standards

Declaration

Ethical Approval

Not Applicable

Competing interests

The authors declare no known competing financial interests

Data Availability

Data shall be made available on request

Conflict of Interest

The authors declare no conflict of interest

Ethical Considerations

Not applicable

Funding

The authors declared no external source of funding.