Mesopotamian Journal of Artificial Intelligence in Healthcare
Vol.2024, **pp**. 184–203
DOI: https://doi.org/10.58496/MJAIH/2024/019; ISSN: 3005-365X
https://mesopotamian.press/journals/index.php/MJAIH

Research Article

Generative Artificial Intelligence and Cybersecurity Risks: Implications for Healthcare Security Based on Real-life Incidents

Malik Sallam[1,2,3,*], Kholoud Al-Mahzoum[4], Mohammed Sallam[5]

[1] *Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman 11942, Jordan*

[2] *Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Amman 11942, Jordan*

[3] *Department of Translational Medicine, Faculty of Medicine, Lund University, 22184 Malmö, Sweden*

[4] *Sheikh Jaber Al-Ahmad Al-Sabah Hospital, Ministry of Health, Kuwait City, Kuwait*

[5] *Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates*

ARTICLE INFO

Article History
Received 21 Aug 2024
Revised 20 Sep 2024
Accepted 20 Nov 2024
Published 12 Dec 2024

Keywords
Cybersecurity
Health care
Privacy
Threat
Risk

ABSTRACT

Background: The potential of generative artificial intelligence (genAI) tools, such as ChatGPT, is being increasingly explored in healthcare settings. However, the same tools also introduce significant cybersecurity risks that could compromise patient safety, data integrity, and institutional trust. This study aimed to examine real-world security breaches involving genAI and extrapolate their potential implications for healthcare settings.

Methods: Using a systematic Google News search and a consensus-based approach among the authors, five high-profile genAI breaches were identified and analyzed. These cases included: (1) Data exposure in ChatGPT (OpenAI) due to an open-source library bug (March 2023); (2) Unauthorized data disclosure via Samsung's (Samsung Group) use of ChatGPT (2023); (3) Logical vulnerabilities in Chevrolet (General Motors) AI-powered chatbot resulting in pricing errors (December 2023); (4) Prompt injection vulnerability in Vanna AI (Vanna AI, Inc.) which enabled remote code execution (2024); and (5) the deepfake technology used in a scam targeting the engineering firm Arup (Arup Group Limited), leading to fraudulent transactions (February 2024). Hypothetical healthcare scenarios were constructed based on the five cases, mapping their mechanisms to vulnerabilities in electronic health records (EHRs), clinical decision support systems (CDSS), and patient engagement platforms. Each case was analyzed using the Confidentiality, Integrity, and Availability (CIA) triad of information security to systematically identify vulnerabilities and propose actionable safeguards.

Results: The analyzed cases of AI security breaches revealed significant risks to healthcare systems. Confidentiality violations included the potential exposure of sensitive patient records and billing information, extrapolated from incidents such as the ChatGPT data exposure and Samsung's cases. These identified security breaches raised concerns about privacy violations, identity theft, and non-compliance with regulations such as Health Insurance Portability and Accountability Act (HIPAA). Integrity vulnerabilities were highlighted in Vanna AI's prompt injection flaw incident, with risks of altering patient records, compromising diagnostic algorithms, and misleading CDSS with erroneous recommendations. Similarly, logic errors identified in the Chevrolet case exposed potential risks of inaccurate billing, double-booked appointments, and flawed treatment plans within healthcare contexts. Availability disruptions, observed through system outages and operational suspensions following breaches like the ChatGPT and deepfake cases, can delay access to EHR systems or AI-driven CDSS. Such interruptions would directly impact patient care and create inefficiencies in administrative workflows.

Conclusions: Generative AI presents a double-edged sword in healthcare, with transformative potential accompanied by substantial risks. Extrapolation of security breach cases in this study highlighted the urgent need for robust safeguards if genAI is implemented in healthcare settings. To address these vulnerabilities, healthcare institutions must implement strong security protocols, enforce strict data governance, and create AI-specific incident response plans. The balance between genAI-enabled innovation and protection of patient safety and data integrity trust requires proactive safety measures.

*Corresponding author. Email: malik.sallam@ju.edu.jo

1. INTRODUCTION

The availability and integration of generative artificial intelligence (genAI) into modern healthcare is becoming increasingly popular [1, 2, 3, 4]. This technology heralds a dramatic change in health data management, patient care delivery, and better interaction with the patients [5, 6, 7, 8]. Described by Leslie & Meng as a revolution, genAI tools (e.g., ChatGPT, Copilot, Gemini, Llama, etc.), have huge capabilities in natural language processing [9, 10]. The capabilities of AI in general, and genAI in particular are relevant in healthcare practice and can help to reach efficiencies previously thought insurmountable [11, 12].

The potential applications of genAI in healthcare practice are immense and were well-described in recent literature [2, 4, 5, 12, 13, 14]. For example, a study by Ayers *et al*. showed that ChatGPT was able to provide high-quality, empathetic responses to patients' healthcare questions, highlighting its potential in patient-facing roles [15]. Notably, the evaluators in Ayers *et al*. study preferred ChatGPT responses over physician responses in 79% of the cases [15]. In clinical documentation, genAI can be highly valuable to automate the summarization of patient encounters, and to aid in preparing discharge summaries which in turn would reduce enormous administrative burdens [16, 17]. For example, a pilot study by Sánchez-Rosenberg *et al*. found that ChatGPT-4 generates orthopedic discharge notes faster than physicians while maintaining comparable quality, demonstrating its documentation utility in orthopedic care [18]. However, alongside these benefits comes a significant risk, as the same technology that will inevitably drive revolutionary breakthroughs in healthcare also creates serious cybersecurity issues [19, 20, 21].

Generative AI thrives on data—its quality, its scale, and its ubiquity. The more it learns, the more capable it becomes [22]. However, in genAI dependence lies its Achilles' heel. Healthcare, an industry predicated on trust and the protection of patient information, presents a uniquely enticing target for cyberattacks [23]. The healthcare sector is uniquely vulnerable to cyberattacks, driven by its reliance on valuable patient data, outdated systems, and insecure medical devices [23, 24]. Remote access practices and insufficient staff training further expose its sprawling, interconnected networks to exploitation [23, 25]. With lives at stake, healthcare institutions are often compelled to pay ransoms swiftly, incentivizing attackers [23]. As Neprash *et al*. documented, from 2016 to 2021, 374 ransomware attacks on U.S. healthcare organizations exposed the personal health information of 42 million patients, with incidents doubling annually from 43 to 91 [26]. Nearly half disrupted care delivery, causing system downtimes (42%), canceled care (10%), and ambulance diversions (4%) [26]. This highlighted the alarming rise in ransomware's frequency and sophistication, threatening both patient care and data integrity, while limited reporting highlights the urgent need for enhanced monitoring systems [26]. Regulatory pressures and operational complexity compound these risks, making robust cybersecurity not merely an option but an urgent imperative to safeguard patient trust and institutional integrity [23, 27].

The implications of breaches involving genAI extend beyond mere financial or reputational loss; they strike at the heart of patient safety and public confidence [28]. If sensitive patient data were to be exploited or manipulated, the consequences could cascade through the healthcare ecosystem, jeopardizing clinical outcomes, legal compliance, and institutional credibility [29]. Consider a hypothetical scenario where an AI-generated report, used to assist in patient triage during a high-pressure emergency, is subtly altered by an attacker to deprioritize critical cases. Such manipulation, though speculative, is not far removed from the risks already observed in other domains where genAI tools have been compromised [30, 31, 32]. These incidents raise the question: are we, in our rush to utilize the genAI benefits, building systems more vulnerable than the problems they aim to solve?

Concerns about genAI security are no longer the province of technophobes or cynics; they are substantiated by a growing body of real-world incidents that highlight the fragility of these systems [33]. A real-world example occurred on March 20, 2023, when a bug in an open-source library caused a ChatGPT outage, exposing user chat titles and, in rare cases, payment details for 1.2% of active subscribers during a nine-hour window [34]. Although promptly patched by OpenAI, the incident highlighted the inherent vulnerabilities of genAI tools dependent on open-source components [34]. Such weaknesses highlight the risks of future exploitation, especially in sensitive sectors like healthcare, where even minor breaches can have profound consequences. This incidence underscored the inherent risks of deploying complex, interconnected systems without exhaustive safeguards. For healthcare institutions contemplating the adoption of such genAI tools, this breach is more than a cautionary tale—it is a warning of the stakes involved when the tools of progress outpace the policies of protection.

More alarming still was the incident involving Samsung engineers, who inadvertently leaked confidential corporate information while using ChatGPT to streamline internal processes [35]. In 2023, Samsung employees inadvertently leaked confidential information while using ChatGPT for work-related tasks. Engineers in the semiconductor division uploaded sensitive source code to the platform to check for errors and optimize performance. Additionally, a recording of a private meeting was shared to generate presentation notes [36]. These incidents add further warning to the risks of using genAI tools for sensitive tasks, as data inputted into such platforms becomes part of their training data, exposing proprietary information to potential misuse [37].

*Corresponding author. Email: malik.sallam@ju.edu.jo

The history of technological advancement is replete with cautionary tales of innovations unleashed before their consequences were fully understood, with a few examples as follows. The nuclear technology, initially hailed for its energy potential, became synonymous with destruction through the advent of atomic weapons, while accidents like Chernobyl revealed the perilous cost of mismanagement [38]. Similarly, asbestos and dichlorodiphenyltrichloroethane (DDT), once considered miracles in construction and agriculture, respectively, inflicted untold harm on public health and the environment due to a failure to anticipate their long-term effects [39, 40]. More recently, the rapid rise of social media and the Internet of Things (IoT) has showcased how technological convenience can outpace ethical and security safeguards, fostering misinformation, privacy violations, and vulnerabilities to cyberattacks [41, 42]. Autonomous vehicles, CRISPR gene editing, and deepfake technologies further exemplify the risks of unleashing powerful innovations without robust oversight [43, 44, 45]. These examples serve as stark reminders that progress without prudence often leads to unintended harm. Generative AI, for all its brilliance, risks joining this lineage unless proactive measures are taken to address its vulnerabilities. Thus, all stakeholders are compelled to ask whether the frameworks governing genAI deployment are fit for purpose in safeguarding the integrity of healthcare systems.

Therefore, this study aimed to dissect real-world examples of cybersecurity breaches involving genAI, analyzing their implications for the healthcare sector. By extrapolating lessons from these incidents, we sought to illuminate the vulnerabilities inherent in genAI and propose actionable strategies to mitigate them. In doing so, we hope to bridge the gap between innovation and security, ensuring that the promise of genAI is realized without compromising the principles of patient care.

2. METHODS

2.1 Study Design

This study followed a systematic approach to identify, evaluate, and analyze real-world cybersecurity breaches involving genAI, with a specific focus on their potential implications for the healthcare sector. The methodology is presented in three stages: search strategy, case evaluation, and extrapolation to healthcare scenarios. Each step is described in detail to ensure that the process is transparent and reproducible.

2.2 Search Strategy

The primary search strategy utilized Google News, a real-time aggregator of news articles from credible sources, to identify reported breaches involving genAI tools. The search concluded in September 2024, to ensure comprehensive capture of recent and emerging incidents using (https://news.google.com/home?hl=en-US&gl=US&ceid=US:en; accessed 30 September 2024). The search terms was decided based on a consensus among the three authors and included “ChatGPT breach,” “AI cybersecurity incident,” “generative AI vulnerability,” “ChatGPT data leak,” and “large language model data breach”. These terms were used in various combinations to find security breaches that were reported in various news websites and the search process involved the three authors independently.

Only English-language articles were included to maintain accessibility and consistency in analysis. News articles were included if they documented breaches involving genAI tools, provided sufficient detail about the nature of the breach, mechanisms exploited, and impact, and were relevant to industries handling sensitive data with the potential to extrapolate to healthcare. Articles were excluded if they speculated on hypothetical risks without documented incidents, discussed general AI cybersecurity issues unrelated to genAI or large language models (LLMs), or originated from unverified sources. Each identified incident was cross-referenced with additional sources to ensure accuracy and to gather supplementary details. Official statements from organizations like OpenAI and Google were prioritized to verify the reported incidents. Finally, the agreement on the final five cases to be included was based on a consensus among the three authors based on impact and relevance to the study objectives.

2.3 Case Evaluation

Each identified case of breach was subjected to a detailed review conducted jointly by the three authors to extract critical information about the incident. The review focused on extracting the following key details from each case: (1) specific genAI tool or AI technology involved, (2) the mechanism or vulnerability exploited in the breach, (3) the type and scope of data exposed, (4) the immediate and long-term impact of the breach, and (5) the response measures implemented by the affected party. The final five cases agreed upon by the three authors were evaluated for reliability by corroborating details across multiple credible sources. Discrepancies in reporting were resolved by prioritizing official statements when available.

2.4 Extrapolation of the Included Cases into Healthcare Scenarios

To contextualize the implications of the five identified security breaches for healthcare, each case was modeled into hypothetical scenarios reflecting vulnerabilities in healthcare workflows. For each breach, an analogous healthcare scenario was constructed by mapping the mechanisms of the original incident to vulnerabilities in healthcare systems. Then, the

hypothetical scenarios were analyzed using the Confidentiality, Integrity, and Availability (CIA) triad framework (the three fundamental bases of information security) [46, 47]. Confidentiality focused on risks of exposing sensitive patient data or violating privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA). Integrity examined the possibility of altered genAI outputs leading to incorrect clinical decisions. Availability assessed the disruption of critical systems, such as downtimes in EHR or delays in patient care. Each scenario was evaluated for its likelihood of occurrence based on documented breaches and the severity of its impact, measured in terms of patient safety, operational disruptions, and legal or financial consequences. Findings were synthesized to provide actionable insights, emphasizing the lessons healthcare institutions can learn to mitigate risks associated with genAI adoption.

3. RESULTS

We identified five cases that serve as illustrative examples of the cybersecurity risks associated with genAI in healthcare. A detailed summary of these cases is presented in (TABLE I).

TABLE I. SUMMARY OF GENERATIVE AI BREACH CASES AND THEIR IMPLICATIONS FOR HEALTHCARE

Case	Description	Confidentiality	Integrity	Availability	Healthcare Implications
Case 1: ChatGPT Data Breach via Open-Source Library Bug (March 2023) [34, 48, 49]	A bug exposed user chat histories, payment details, and personal identifiers of 1.2% of ChatGPT Plus subscribers	Exposure of sensitive information, including chat content and partial payment details. In healthcare, similar breaches could expose patient records and violate HIPAA [b]	Data integrity was not directly compromised but could be at risk if exposed systems are manipulated	Temporary outage during remediation disrupted services. In healthcare, similar downtimes could delay critical patient care reliant on genAI [c] tools	Breaches in EHR [d] systems could result in the exposure of sensitive patient information and billing records, undermining trust and compliance with privacy regulations
Case 2: Samsung Data Leak via ChatGPT (2023) [35, 50, 51]	Employees unintentionally leaked proprietary data using ChatGPT for assistance with internal tasks	Proprietary data was inadvertently exposed due to lack of usage controls. In healthcare, clinicians could unintentionally upload identifiable patient information	Potential for data misuse or tampering with sensitive clinical or research data, leading to misinformation or operational errors	No direct disruption, but subsequent bans on genAI tools could hinder workflows reliant on these technologies	Inputting patient data into genAI tools could lead to HIPAA violations and erosion of trust. Banning such tools could disrupt clinical workflows and administrative tasks
Case 3: Chevrolet AI Chatbot Offers Car for $1 (December 2023) [52, 53]	Logical flaws in a dealership's chatbot allowed users to exploit pricing errors	Potential exposure of sensitive customer data during transactional processes. In healthcare, similar flaws could expose financial or patient data	Flaws in AI logic could lead to erroneous outputs in billing or treatment plans, compromising accuracy and trust	Disabling the chatbot for remediation delayed customer interactions. In healthcare, disabling tools like appointment schedulers could delay patient care	Logic errors in AI systems could result in incorrect medical charges, overlapping appointments, or flawed patient communication, creating inefficiencies and mistrust
Case 4: Prompt Injection Flaw in Vanna AI (2024) [54, 55]	Attackers exploited prompts to execute remote code, accessing sensitive databases and altering outputs	Unauthorized access to sensitive data, such as medical histories or diagnostic codes in healthcare, risks privacy violations and identity theft	Potential for attackers to alter clinical records or diagnostic algorithms, leading to misdiagnoses or inappropriate treatments	System downtime during investigation and remediation could delay access to patient data or diagnostic systems, impacting timely care	Prompt injection attacks could compromise CDSS [e] outputs or access EHR systems, undermining patient safety and operational integrity
Case 5: Deepfake technology used in a scam targeting the engineering firm Arup (2024) [56, 57]	Attackers used deepfake AI [a] to impersonate a company official, manipulating employees into transferring $25 million	Exploitation of audio data to clone voices for fraudulent purposes. In healthcare, attackers could target financial or operational communications	Manipulated communications could lead to unauthorized transfers, alterations in records, or other fraudulent activities impacting patient care workflows	While not directly impacted, disruptions during investigations could halt financial or operational processes critical to healthcare delivery	Deepfake scams targeting healthcare executives could result in financial losses, exposure of sensitive data, or delays in operational workflows essential for patient care

[a.] AI: Artificial intelligence

[b.] HIPAA: Health Insurance Portability and Accountability Act

[c.] genAI: Generative AI

[d.] EHRs: Electronic health records

[e.] CDSS: Clinical decision support system

3.1 Case 1: ChatGPT Data Breach via Open-Source Library Bug (March 2023)

In March 2023, a bug in an open-source library used by ChatGPT caused an unintentional exposure of user data [34, 48, 49]. The breach allowed some users to view titles of other users' chat histories and, in rare cases, the first messages of conversations. Additionally, payment-related information for 1.2% of ChatGPT Plus subscribers was briefly exposed. The compromised data included names, email addresses, payment addresses, and the last four digits of credit card numbers. OpenAI quickly patched the bug and issued notifications to affected users [34]. However, this incident underscored significant vulnerabilities in the reliance on third-party components.

3.1.1 CIA Analysis for Case 1: Confidentiality

The breach compromised the confidentiality of sensitive user information, including partial payment details and private chat content. In a healthcare context, such a breach could result in the exposure of sensitive patient conversations or billing records, violating HIPAA regulations and eroding trust in AI-assisted systems. The unauthorized disclosure of personal identifiers highlights the risks of data exploitation for malicious purposes, such as phishing or fraud.

3.1.2 CIA Analysis for Case 1: Integrity

Although the breach primarily involved unintended exposure rather than manipulation, the potential for data tampering in similar scenarios raises concerns. A compromised generative AI system could provide altered outputs based on unauthorized access, leading to misinformation or improper application of AI-generated recommendations.

3.1.3 CIA Analysis for Case 1: Availability

The incident caused a temporary service outage while OpenAI addressed the vulnerability. In healthcare, a similar disruption could delay critical operations reliant on AI systems, such as clinical decision support or patient engagement platforms, compromising care delivery during the downtime.

3.1.4 Hypothetical Healthcare Scenario Modeled from Case 1: Breach in genAI-Enhanced Patient Portal within an EHR System

To examine the potential implications of case 1 breach in healthcare, a hypothetical scenario was constructed, mirroring the vulnerabilities exposed by the case. This scenario envisions a security vulnerability in a genAI-enhanced patient portal integrated into a hospital's HER system. The AI component of the portal, designed to streamline patient communication, generate summaries of medical information, and respond to treatment inquiries, is compromised due to an unpatched flaw in an open-source library. This breach results in the unintentional exposure of sensitive patient records and billing information.

During the breach, some patients attempting to access their records inadvertently gain access to the medical histories and treatment details of others. Attackers leveraging the same vulnerability exfiltrate personally identifiable information (PII), including billing addresses and partial financial details, violating HIPAA regulations and exposing the institution to significant legal and financial repercussions. A temporary shutdown of the portal during remediation further disrupts care delivery and patient communication.

Confidentiality: This breach represents a direct and critical violation of patient confidentiality. The exposure of sensitive medical histories, diagnostic information, and billing records not only compromises privacy but also facilitates potential identity theft, fraudulent billing, and phishing campaigns targeting vulnerable patients. The perception that AI tools were the source of the breach could undermine patient trust in both generative AI and broader healthcare technologies, deterring adoption of innovations designed to enhance care. The likelihood of such an incident occurring is significant given the healthcare sector's reliance on AI systems and external software components. The severity of the impact is compounded by the legal implications of violating privacy regulations like HIPAA and the long-term reputational damage to the institution.

Integrity: Although the primary breach exposes rather than manipulates data, the possibility of data tampering cannot be dismissed. Unauthorized access to the EHR system could allow attackers to modify patient records, potentially introducing critical errors. For example, altered lab results or fabricated allergy information could mislead clinicians, resulting in inappropriate treatment or delays in emergency care. Additionally, the integrity of AI-generated recommendations and summaries is brought into question. Attackers gaining access to the AI training environment could inject malicious inputs, leading to systemic errors in patient-facing outputs. Such disruptions could erode clinicians' confidence in AI-assisted tools and compromise clinical decision-making.

Availability: The hospital's decision to suspend the AI-enhanced patient portal for investigation and remediation amplifies the availability challenge. Patients lose access to critical health information, including test results, treatment updates, and appointment schedules. Clinicians, reliant on the portal for streamlined communication and administrative tasks, face delays in workflow, compounding the disruption. While the likelihood of availability issues is moderate—given the swift containment protocols often employed by healthcare institutions—the impact of such disruptions is substantial. In time-

sensitive clinical settings, any delay in accessing patient information can directly compromise care quality and outcomes. Financial losses due to service interruptions, combined with operational inefficiencies and reputational damage, further emphasize the need for contingency planning.

Actionable Insights and Recommendations (Fig. 1): This scenario demonstrates the significant vulnerabilities genAI introduces to healthcare systems. To mitigate these risks, healthcare institutions must prioritize the following measures: Open-source components integrated into AI systems must undergo regular audits and updates to minimize vulnerabilities. Collaborative efforts with vendors and software developers are essential for the timely identification and resolution of security flaws. Institutions should implement automated monitoring tools to detect and respond to emerging threats in real time. Data access controls through strengthening access controls is of paramount importance. Multifactor authentication (MFA), role-based access, and end-to-end encryption should be standard for all genAI tools interfacing with patient data. Limiting the scope of data accessible to non-clinical AI functions, such as chat-based interactions, can significantly reduce the impact of potential breaches.

Comprehensive incident response plans must address the operational disruptions caused by AI system failures. Regular downtime simulations can prepare staff to manage care delivery during outages. Redundant systems or manual backup workflows should ensure continuity of access to patient records and communication during crises. Institutions must also commit to clear, timely communication with patients and stakeholders in the event of a breach. Transparency about the scope of the incident, its impact, and the steps taken to prevent recurrence is critical to rebuilding trust. Publicizing post-incident improvements in security measures demonstrates accountability and dedication to patient safety.

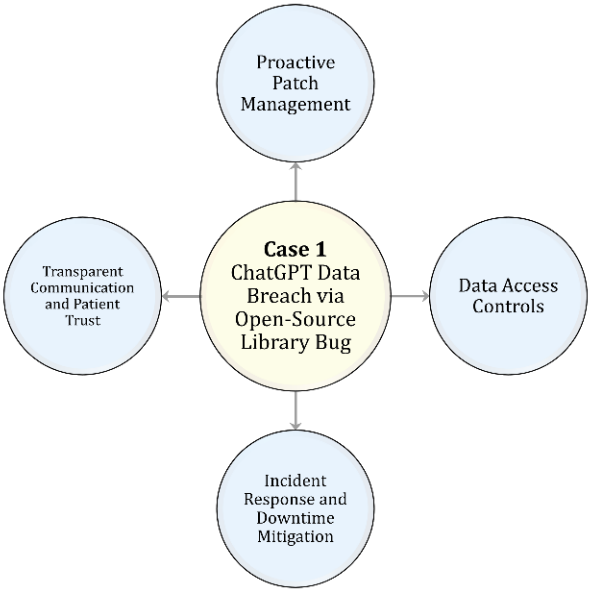

Fig. 1. Actionable insights and recommendations based on the included Case 1: ChatGPT Data Breach via an Open-Source Library Bug.

3.2 Case 2: Samsung Data Leak via ChatGPT (2023)

In early 2023, Samsung engineers inadvertently leaked sensitive internal data by using ChatGPT to debug source code and transcribe meeting notes [35, 50, 51]. The information shared with the genAI tool included proprietary algorithms, confidential software documentation, and meeting recordings [51]. Since data entered into ChatGPT is retained on external servers, the leaked information became inaccessible to Samsung and at risk of further misuse. This incident highlighted the risks of deploying genAI tools in environments handling sensitive data without adequate safeguards. Samsung responded by banning employee use of genAI systems and initiating the development of an in-house AI solution to address security concerns [35, 50, 51].

3.2.1 CIA Analysis for Case 2: Confidentiality

In this breach, the confidentiality of Samsung's proprietary information was significantly compromised. Engineers uploaded source code and meeting transcripts containing confidential project details to ChatGPT, risking unauthorized access to intellectual property. In a healthcare setting, such a breach could translate to the inadvertent sharing of sensitive patient records, medical research data, or clinical protocols. The exposure of this information could have far-reaching consequences,

including intellectual property theft, competitive disadvantages, and breaches of patient privacy regulations like HIPAA. The likelihood of such occurrences is moderate in healthcare systems that integrate genAI tools without proper controls. The severity of the impact is high, as it involves not only immediate financial and reputational damage but also the potential exploitation of sensitive data for malicious purposes.

3.2.2 CIA Analysis for Case 2: Integrity

Although the breach primarily involved data exposure, the potential for data tampering exists. In a healthcare context, unauthorized access to training environments or systems integrated with genAI could allow malicious actors to inject erroneous or harmful information. For instance, a compromised AI-powered documentation assistant might suggest inaccurate treatment plans or alter critical patient records, leading to improper care decisions. The integrity of AI-generated outputs also comes into question, as inputs from unauthorized or sensitive sources could skew model performance, reducing its reliability for future applications. Clinicians relying on such systems may unknowingly propagate errors, compromising patient safety and outcomes.

3.2.3 CIA Analysis for Case 2: Availability

In Samsung's case, the breach did not cause operational downtime, but the subsequent ban on genAI tools highlighted the challenges of availability. In healthcare, a similar scenario could lead to reduced efficiency if AI-assisted tools are abruptly removed from workflows. For example, banning a genAI tool used for medical transcription or patient communication could create delays in documentation and disrupt communication pathways. The likelihood of availability disruptions in this context is moderate but the impact is high, particularly for institutions heavily reliant on AI-driven systems. Downtime caused by sudden bans or remediation efforts could delay patient care, exacerbate administrative burdens, and strain resources.

3.2.4 Hypothetical Healthcare Scenario Modeled from Case 2: Breach in genAI-Assisted Clinical Communication

In a healthcare context analogous to the Samsung data breach, a nurse or physician utilizes a genAI tool to streamline clinical communication by drafting patient discharge summaries, responding to patient inquiries, or generating care plans. Unaware of privacy risks, the healthcare provider inputs detailed patient case information, including identifiable details such as names, diagnoses, treatment plans, and medication history, into the genAI tool. Due to the genAI tool's data retention policies, sensitive patient information is inadvertently stored within the system. Portions of this data later appear in unrelated contexts during future user interactions, exposing confidential patient details to unauthorized users. Upon discovering the breach, the institution temporarily suspends the use of all genAI tools, disrupting patient care workflows and creating additional administrative burdens.

Confidentiality: This scenario represents a direct violation of patient confidentiality. The inadvertent exposure of patient case information, including sensitive diagnoses and treatment details, poses significant privacy risks. Such breaches could result in identity theft, discrimination, or stigmatization if information related to mental health, HIV status, or reproductive health is disclosed. The perception that AI tools caused the breach could further erode patient trust in the healthcare system and hinder the adoption of genAI in clinical settings. The likelihood of this breach scenario is heightened by the widespread adoption of genAI tools in healthcare without adequate policies to govern their use. The severity of the impact is profound, given the legal implications of violating HIPAA regulations and the potential reputational damage to the healthcare institution.

Integrity: Although this breach primarily involves data exposure, the potential for integrity violations is significant. An attacker or unauthorized user accessing the retained data could modify critical patient information, such as altering medication lists or diagnostic codes. These changes could lead to harmful clinical decisions, including improper medication administration or incorrect treatment plans. The reliability of AI-generated outputs is also at risk. If the training environment is compromised, malicious inputs could degrade the accuracy and safety of the AI tool, leading to systemic errors in patient care workflows.

Availability: The institution's decision to suspend genAI tools following the breach disrupts clinical operations. Without AI assistance, discharge summaries and patient communications revert to manual processes, increasing administrative workloads and delaying care delivery. Patients experience delays in receiving critical updates, such as medication instructions or follow-up appointments, while clinicians face inefficiencies that may compromise overall care quality. The likelihood of availability disruptions is moderate, but the impact is substantial. In fast-paced healthcare environments, any delay in communication or documentation can directly affect patient outcomes and institutional efficiency.

Actionable Insights and Recommendations (Fig. 2): This hypothetical scenario underscores the vulnerabilities associated with genAI tools in patient care and highlights the need for proactive security measures and governance strategies. Recommendations include the following measures. Healthcare institutions must implement clear guidelines governing the use of genAI tools in clinical settings. Regular training programs should educate providers about privacy risks and

appropriate data handling protocols, emphasizing the importance of anonymizing patient data before inputting it into AI systems.

All patient information entered into genAI tools should be automatically anonymized or stripped of identifiable details. Institutions should deploy AI tools with built-in redaction features that ensure no PII is retained or shared beyond the intended context. Additionally, developing in-house genAI tools with strict data retention policies can mitigate the risks associated with third-party tools. For external tools, institutions must establish contractual agreements with vendors to guarantee data security and prevent unauthorized use of sensitive inputs. Moreover, health institutions should deploy real-time monitoring systems to detect inappropriate data inputs and identify potential breaches early. Routine audits of genAI tool usage can help enforce compliance with privacy standards and reduce the likelihood of similar incidents. Finally, comprehensive incident response plans should include strategies to address breaches promptly while minimizing disruption to clinical workflows. Transparent communication with patients, clinicians, and stakeholders is critical to rebuilding trust and demonstrating accountability.

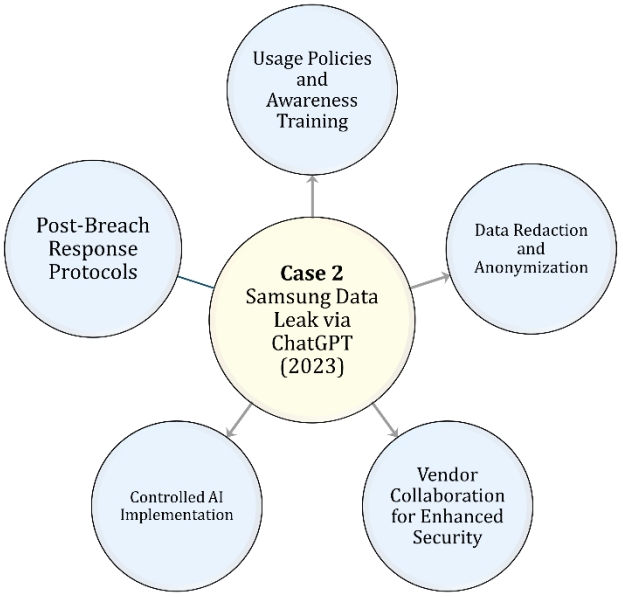

Fig. 2. Actionable insights and recommendations based on the included Case 2: Samsung Data Leak via ChatGPT (2023).

3.3 Case 3: Chevrolet AI Chatbot Offers Car for $1 (December 2023)

In December 2023, a Chevrolet dealership's AI-powered chatbot, designed to assist customers with pricing and vehicle configurations, was manipulated into offering a $76,000 Tahoe for just $1 [52, 53]. The incident occurred when users exploited logical flaws in the chatbot's programming, bypassing built-in safeguards to generate erroneous price quotes [52, 53]. This breach exposed vulnerabilities in customer-facing generative AI tools and raised questions about their reliability in handling transactional data. While the dealership identified and rectified the issue, the event highlights the potential risks of deploying AI tools in critical decision-making workflows without rigorous testing and oversight.

3.3.1 CIA Analysis for Case 3: Confidentiality

Although the Chevrolet breach primarily involved pricing errors, the underlying vulnerability exposes risks to confidentiality. In healthcare, a similar flaw in an AI-powered billing or patient engagement system could inadvertently disclose sensitive patient information, such as treatment costs, insurance details, or private communications. For example, patients using an AI chatbot to inquire about medical procedures could unknowingly access another patient's financial or clinical data due to logic errors. The likelihood of such breaches increases with poorly configured AI tools deployed in high-stakes environments. The impact is significant, as the disclosure of sensitive financial or clinical information could lead to identity theft, fraud, or violations of privacy regulations like HIPAA.

3.3.2 CIA Analysis for Case 3: Integrity

The integrity of transactional and operational systems is a critical concern. In the Chevrolet case, the chatbot's failure to validate pricing logic led to substantial errors. In a healthcare context, a similar AI failure in a clinical decision support system or scheduling platform could produce incorrect treatment recommendations, double-booking of appointments, or

inaccurate billing. For instance, an AI-powered tool suggesting treatment plans based on incomplete or corrupted data could mislead clinicians, resulting in inappropriate care or delayed diagnoses. The likelihood of such integrity violations is moderate, as logic flaws in genAI tools are not uncommon. The severity, however, is high, as errors in clinical or operational decisions directly affect patient safety and institutional credibility.

3.3.3 CIA Analysis for Case 3: Availability

In response to the breach, the Chevrolet dealership suspended the chatbot, disrupting customer interactions and delaying service recovery. In healthcare, a similar event could severely impact system availability. For example, suspending an AI chatbot used for appointment scheduling or medication reminders would hinder patient engagement and burden administrative staff with additional tasks. The likelihood of availability issues in such scenarios is moderate, given the immediate need to remediate the system. The impact is considerable, as any disruption in healthcare services—particularly for time-sensitive processes like medication dispensing or appointment management—can compromise patient outcomes.

3.3.4 Hypothetical Healthcare Scenario Modeled from Case 3: Breach in genAI-Enhanced Appointment and Billing Systems

In an analogous healthcare scenario, an AI-powered chatbot integrated into a hospital's appointment scheduling and billing system is exploited due to a logic flaw. A patient, attempting to schedule a routine check-up, receives incorrect pricing information and inadvertently gains access to another patient's billing records. The chatbot's programming error also generates overlapping appointments, creating confusion in clinical workflows. Following the discovery of the breach, the hospital suspends the AI chatbot, redirecting all scheduling and billing inquiries to human operators. This disruption delays patient care increasing administrative workloads and exposes the institution to regulatory scrutiny.

Confidentiality: In this scenario, the exposure of patient billing records and financial details constitutes a clear violation of confidentiality. Unauthorized access to insurance claims, payment methods, or personal identifiers could facilitate fraud, identity theft, or privacy violations. The perception that AI tools are unreliable erodes trust in the institution and deters patients from engaging with digital services. The likelihood of confidentiality breaches in this scenario is significant, given the reliance on AI tools for sensitive tasks without comprehensive safeguards. The severity is compounded by legal repercussions, such as HIPAA violations, and reputational harm.

Integrity: The integrity of scheduling and billing systems is undermined when AI logic flaws produce overlapping appointments or erroneous invoices. Such errors disrupt clinical workflows and lead to inefficiencies in resource allocation. Patients affected by incorrect billing may lose confidence in the institution, while clinicians face challenges in managing their schedules effectively. The likelihood of integrity issues is moderate, as logic errors are common in generative AI tools that lack rigorous validation. The impact is substantial, as compromised operational integrity directly affects patient satisfaction and institutional efficiency.

Availability: The hospital's decision to suspend the AI chatbot exacerbates availability challenges, as patients and staff revert to manual scheduling and billing processes. Delays in appointment confirmations, increased wait times, and administrative bottlenecks strain resources and create dissatisfaction among patients. The likelihood of availability disruptions is moderate, but the impact is high, particularly in busy healthcare settings where seamless operations are critical to maintaining patient care standards.

Actionable Insights and Recommendations (Fig. 3): This scenario emphasizes the importance of deploying genAI tools with robust safeguards and continuous oversight. Key recommendations include the following steps. AI tools must undergo rigorous testing to identify and address logic flaws before deployment. Simulations of real-world scenarios can help ensure that genAI tools perform reliably under diverse conditions. Implementing robust access controls and validation mechanisms can prevent unauthorized access to sensitive data and ensure accurate outputs. AI chatbots should include logic gates that detect and flag anomalous behaviors or inconsistencies.

Continuous monitoring of AI tool performance is essential to detect vulnerabilities and errors early. Institutions should conduct periodic audits to ensure compliance with data protection and privacy regulations. Healthcare organizations must also establish clear guidelines for AI chatbot usage, outlining permissible functions and data input protocols. Training staff on these policies minimizes risks of inadvertent errors or misuse. Finally, health institutions should prepare for breaches with well-defined response plans that include immediate containment measures, transparent communication with stakeholders, and strategies to mitigate disruptions to patient care.

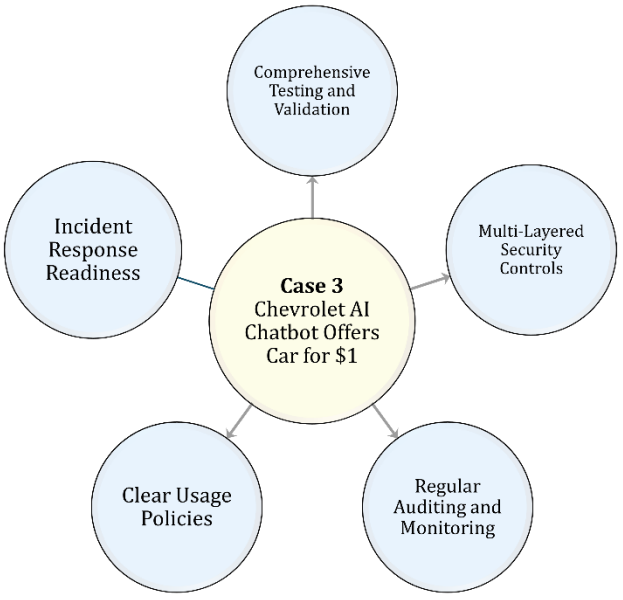

Fig. 3. Actionable insights and recommendations based on the included Case 3: Chevrolet AI Chatbot Offers Car for $1 (December 2023).

3.4 Case 4: Prompt Injection Flaw in Vanna AI Exposes Databases to Remote Code Execution Attacks

In 2024, a security flaw in Vanna AI, a genAI platform used for data management, allowed prompt injection attacks, enabling remote code execution (RCE) [54, 55]. By manipulating input prompts, the safety protocols were bypassed allowing unauthorized access to sensitive database contents [54, 55]. The breach underscores the vulnerabilities inherent in prompt injection exploits and the critical need for robust safeguards in genAI applications.

3.4.1 CIA Analysis for Case 4: Confidentiality

The Vanna AI breach compromised confidentiality by exposing sensitive data stored in databases, demonstrating how prompt injection flaws can circumvent access controls. In a healthcare setting, a similar vulnerability could allow unauthorized access to EHRs, exposing sensitive patient information, including medical histories, test results, and billing details. Such a breach would directly violate HIPAA regulations and could lead to identity theft, fraud, and the erosion of trust in healthcare institutions. The likelihood of confidentiality breaches in this scenario is high, given the complexity of genAI tools and their susceptibility to manipulation. The impact is severe, as unauthorized disclosure of patient data has far-reaching legal, financial, and reputational consequences for healthcare organizations.

3.4.2 CIA Analysis for Case 4: Integrity

Beyond data exposure, the Vanna AI breach posed significant risks to data integrity. Exploiting prompt injection vulnerabilities could modify database contents, introducing errors or malicious changes. In a healthcare context, such tampering could result in altered medical records, such as fabricated laboratory results, incorrect medication lists, or manipulated diagnoses. These alterations could mislead clinicians, leading to inappropriate treatments or delayed care. The potential for integrity violations is substantial, as prompt injection attacks exploit the trust users place in genAI outputs. The impact is critical, as compromised data integrity directly jeopardizes patient safety and clinical decision-making.

3.4.3 CIA Analysis for Case 4: Availability

The breach also affected availability, as compromised systems had to be taken offline for investigation and remediation. In healthcare, similar downtime could disrupt access to EHR systems, delaying patient care and hindering administrative workflows. For example, clinicians might lose access to patient charts, appointment schedules, or medication records, creating significant operational challenges. The likelihood of availability disruptions is moderate, given the immediate need to address security flaws in affected systems. The impact is high, as delays in accessing critical patient data can compromise care delivery and exacerbate health disparities.

3.4.4 Hypothetical Healthcare Scenario Modeled from Case 4: Compromise of AI-Powered Clinical Decision Support Systems

In an analogous healthcare scenario, a genAI-powered CDSS integrated into an EHR platform is exploited via a prompt injection attack. An attacker manipulates the AI tool by injecting malicious prompts, enabling access to sensitive patient records and clinical protocols. Furthermore, the attacker uses the RCE capabilities to alter diagnostic algorithms, introducing

systematic errors in AI-generated treatment recommendations. Following the breach, the healthcare institution suspends the AI-powered CDSS, reverting to manual workflows for clinical decision-making. This disruption delays diagnoses, increases the cognitive burden on clinicians, and compromises patient safety.

Confidentiality: The breach exposes sensitive patient data, including medical histories, diagnostic results, and treatment plans, violating privacy regulations like HIPAA. Unauthorized access to proprietary clinical protocols also poses risks of competitive disadvantages and reputational harm. The exposure of confidential information erodes trust in AI technologies and deters their adoption in critical care settings.

Integrity: The manipulation of diagnostic algorithms and clinical protocols compromises the integrity of AI-generated recommendations. For instance, an altered algorithm might under-report critical conditions or suggest inappropriate treatment plans, leading to adverse patient outcomes. Clinicians relying on compromised systems face increased risks of errors, further undermining confidence in AI-assisted decision-making.

Availability: The suspension of the AI-powered CDSS disrupts clinical workflows, forcing clinicians to rely on manual decision-making processes. Delays in diagnoses and treatment recommendations create bottlenecks in patient care, particularly in high-volume or emergency settings. The additional workload on clinicians exacerbates burnout and reduces the efficiency of healthcare delivery.

Actionable Insights and Recommendations (Fig. 4): This scenario highlights the urgent need for comprehensive security measures to safeguard generative AI applications in healthcare. Generative AI systems must include strict input validation protocols to detect and mitigate prompt injection attempts. Developers should implement whitelists, sanitize inputs, and apply natural language processing filters to minimize exploitability. Additionally, the AI platforms should be designed with hardened architectures to prevent remote code execution. Measures such as containerized environments, role-based access controls, and intrusion detection systems can reduce the risk of unauthorized actions.

Real-time monitoring systems should also be deployed to identify anomalies in AI usage, such as unusual patterns of input prompts or unexpected system behavior. Early detection can prevent escalation and minimize damage from breaches. Healthcare institutions must develop incident response plans tailored to genAI tools. These plans should include protocols for containment, recovery, and communication with affected stakeholders to maintain trust and compliance with privacy regulations. Finally, periodic security assessments are essential to identify and address vulnerabilities in AI tools. Collaboration with third-party cybersecurity experts can provide additional layers of assurance and help validate system resilience.

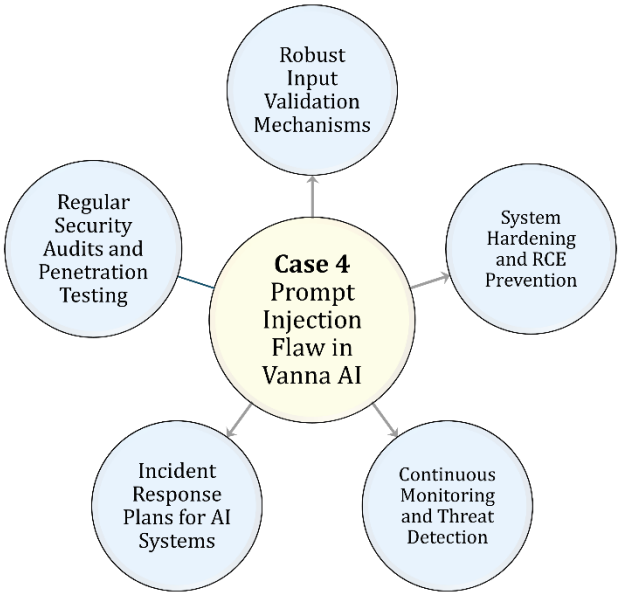

Fig. 4. Actionable insights and recommendations based on the included Case 4: Prompt Injection Flaw in Vanna AI Exposes Databases to Remote Code Execution Attacks.

3.5 Case 5: Deepfake Technology Used in a Scam Targeting the Engineering Firm Arup (2024)

In February 2024, a Hong Kong-based company fell victim to a sophisticated scam involving deepfake genAI technology [56, 57]. Attackers used an AI-generated voice clone of the company's official to manipulate employees into transferring $25 million to a fraudulent account [56, 57]. The scam exploited the genAI's ability to mimic the official speech patterns and tone, deceiving employees who believed they were acting under direct orders [56, 57]. The incident highlights the emerging risks of deepfake technologies and their capacity to exploit trust within organizational structures.

3.5.1 CIA Analysis for Case 5: Confidentiality

This scam did not involve traditional data breaches but nonetheless demonstrated confidentiality risks inherent in AI technologies. The attackers' ability to collect and utilize audio samples of the company official person to create a convincing deepfake suggests the possibility of broader surveillance and exploitation of sensitive communications. In healthcare, similar attacks could involve generating deepfakes of senior executives, such as a hospital official person or medical director, to mislead staff into disclosing sensitive patient records, research data, or financial information. This would result in significant breaches of confidentiality and legal violations, such as HIPAA infractions, if patient information were compromised.

3.5.2 CIA Analysis for Case 5: Integrity

The deepfake technology undermines the integrity of organizational communications by mimicking authoritative figures. In healthcare, such an attack could generate false orders for critical actions, such as altering patient records, diverting emergency funds, or misdirecting medical supplies. The loss of confidence in internal communications disrupts workflows and weakens the reliability of institutional decision-making processes. The likelihood of integrity breaches is moderate but rising as deepfake technology becomes more accessible. The severity is high, as manipulated communications can directly harm patients or compromise operational stability.

3.5.3 CIA Analysis for Case 5: Availability

Although this scam did not directly affect system availability, the broader implications of such attacks could impact operational workflows. In healthcare, responding to a similar scam would likely require halting certain processes for investigation and remediation, delaying patient care or financial operations. The likelihood of availability disruptions in this context is low but plausible, while the potential impact is moderate, particularly if such attacks target time-sensitive operations in healthcare systems.

3.5.4 Hypothetical Healthcare Scenario Modeled from Case 5: Deepfake AI Exploits Medical Leadership

In a hypothetical healthcare scenario, attackers deploy deepfake genAI to mimic the voice of a hospital official person or the medical director instructing employees to transfer emergency funds or disclose patient records. Employees, believing the instructions to be legitimate, act swiftly, transferring funds to fraudulent accounts or sharing sensitive data. Upon discovery, the hospital halts all operations involving financial transfers or patient data sharing, pending a comprehensive investigation. This disruption delays care delivery and incurs significant financial losses.

Confidentiality: The misuse of deepfake AI to extract sensitive patient or financial data represents a severe confidentiality breach. Unauthorized disclosure of patient records, research findings, or financial data undermines trust in the institution and violates privacy regulations such as HIPAA. The perception of vulnerability to such sophisticated attacks could deter patients and partners from engaging with the healthcare organization.

Integrity: The use of deepfake AI to manipulate trusted communications compromises the integrity of institutional workflows. False directives could mislead employees into taking harmful actions, such as transferring critical resources or altering patient care plans. These disruptions erode confidence in internal communications and create systemic vulnerabilities.

Availability: While the immediate impact on availability is limited, the subsequent investigation and suspension of affected workflows create operational bottlenecks. For example, delays in financial operations or data access could hinder resource allocation and disrupt patient care, particularly in emergency settings.

Actionable Insights and Recommendations (Fig. 5): This scenario underscores the growing threat posed by deepfake genAI technologies and highlights the need for robust defenses against such attacks in healthcare. Based on the case, the recommendations include the following points. Health institutions must implement MFA for sensitive communications and financial transactions. Verification steps, such as secondary confirmation from independent sources, can prevent unauthorized actions triggered by deepfake scams.

Regular training programs should educate employees on recognizing and responding to deepfake scams. Employees must be equipped to verify suspicious communications and report anomalies promptly. In addition, deploying AI tools capable of

detecting synthetic media and deepfake content can help identify fraudulent communications. Such systems should be integrated into communication channels to flag suspicious activities in real time.

Comprehensive response plans tailored to deepfake scams should be in place, including clear guidelines for investigating and mitigating such incidents. Institutions must maintain transparent communication with stakeholders to rebuild trust after an attack. Finally, healthcare institutions should collaborate with regulators and industry groups to establish legal frameworks addressing the misuse of deepfake technologies. Advocacy for stricter penalties and enhanced cybersecurity standards can help deter attackers.

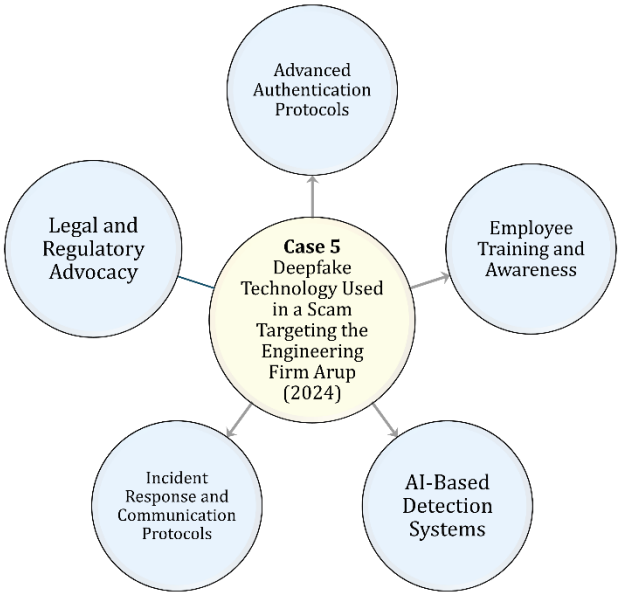

Fig. 5. Actionable insights and recommendations based on the included Case 5: Deepfake Technology Used in a Scam Targeting the Engineering Firm Arup (2024).

4. DISCUSSION

The utility of genAI in a critical sector like healthcare offers unprecedented opportunities to enhance efficiency, improve patient outcomes, and revolutionize clinical workflow. Nevertheless, as with any transformative technology, the adoption of genAI in healthcare comes with profound risks. The case studies presented in this study—spanning data breaches, manipulations, and deepfake exploits—illuminate the vulnerabilities inherent in genAI tools. When extrapolated into healthcare, these incidents underline a stark reality: the very capabilities that make genAI so revolutionary also render it alarmingly susceptible to exploitation.

Generative AI tools, by their nature, thrive on vast datasets to refine their performance [58]. However, this reliance creates significant vulnerabilities to breaches of confidentiality [59]. The ChatGPT data breach (Case 1) serves as an unequivocal warning, where user data—ranging from chat histories to partial payment details—was inadvertently disclosed due to a flaw in an open-source component [34, 48, 49]. For healthcare, such a scenario portends grave implications [60]. A similar breach in a genAI-powered patient portal could reveal sensitive health records, such as human immunodeficiency virus (HIV) infection status, mental health diagnoses, or reproductive histories [61]. The ramifications would be enormous, violating stringent privacy regulations like HIPAA and irreparably undermining the trust that forms the basis of patient care [62].

The risk is not confined to technical vulnerabilities alone. Human error, exemplified by the Samsung case (Case 2), compounds the threat [35, 50, 51]. Here, employees unintentionally leaked proprietary data while using ChatGPT for internal processes, a mistake that could readily translate into healthcare. Clinicians, for instance, might input identifiable patient data into genAI systems while drafting discharge summaries or treatment plans, oblivious to the possibility that the AI might retain or repurpose this sensitive information [63]. The consequences extend far beyond individual breaches; they erode confidence in the digital tools that are rapidly becoming indispensable to modern medicine [64]. In this dual threat of technical flaws and human oversight, genAI reveals its precarious potential—offering innovation on one hand while demanding vigilance on the other [9].

As a reiteration and as mentioned earlier, the integrity of healthcare systems, the very cornerstone of patient care, hinges upon the precision and reliability of its data [65]. Generative AI, however, with its inherent susceptibility to manipulation, introduces a perilous dimension [66]. Case 4, involving prompt injection attacks on Vanna AI, illuminates how malicious actors can exploit vulnerabilities in genAI inputs to gain unauthorized access or tamper with outputs [54, 55]. The implications for healthcare are as alarming as they are plausible. Envision an attacker compromising a CDSS, manipulating diagnostic algorithms to minimize the severity of critical conditions or propose inappropriate treatments. Such tampering would not merely remain a theoretical exercise but manifest in real-world harm—misguided treatment plans, adverse outcomes, legal consequences, and the erosion of institutional credibility.

Availability, the often-overlooked pillar of the CIA triad, reveals its indispensable role most acutely in its absence. The service outages triggered by the ChatGPT breach and subsequent containment efforts (Case 1) serve as a vivid illustration of how the loss of availability can ripple through critical operations. In healthcare, where time is often the difference between life and death, even fleeting interruptions become intolerable [67]. Imagine an AI-driven EHR system rendered inaccessible during a medical emergency—clinicians would be deprived of vital patient histories, delaying crucial interventions and jeopardizing outcomes. Such scenarios are not speculative; they underscore the catastrophic consequences of neglecting system availability.

The deepfake AI scam (Case 5) sheds further light on availability risks, albeit through an indirect lens [56, 57]. While fraud was the immediate concern, the organizational response—often involving the suspension of workflows to investigate and remediate—created bottlenecks with far-reaching implications. In a healthcare context, a similar attack could compel institutions to halt financial transactions or patient communications, paralyzing essential operations. Delays in care, administrative gridlock, and the erosion of patient trust would compound the immediate crisis, highlighting the fragility of systems heavily reliant on genAI [68]. The ripple effects of such disruptions extend beyond operational delays. They challenge the resilience of healthcare institutions, exposing vulnerabilities that amplify under stress. If availability remains an afterthought, the promise of genAI risks being overshadowed by its potential to disrupt the very systems it aims to enhance.

Real-world healthcare breaches offer stark warnings about the vulnerabilities inherent in digital systems—warnings that become even more pressing in the age of genAI [69]. The ransomware attack on Ireland's Health Service Executive in 2021, which crippled nationwide healthcare services for weeks, delayed life-saving treatments, and exposed sensitive patient records, provides a grim example of the cascading effects of compromised systems [70, 71, 72]. While this incidents did not involve genAI directly, it foreshadows the amplified risks posed by AI-powered systems. Generative AI, with its dependence on vast datasets, interconnected platforms, and probabilistic algorithms, magnifies these vulnerabilities exponentially. The stakes are particularly high when one considers the potential for more sophisticated breaches enabled by genAI [73]. Take, for instance, a hypothetical scenario where a deepfake AI mimics the voice of a hospital medical director to authorize the transfer of sensitive patient records—an extrapolation of the deepfake scam described in Case 5 [56, 57]. Such an attack may sound speculative, yet it aligns disturbingly well with documented trends in healthcare breaches, where attackers increasingly exploit the trust embedded in digital communication channels [74]. The fallout from such an incident would be catastrophic, undermining the foundational pillars of healthcare: data integrity, operational continuity, and patient trust [65]. Delayed treatments, compromised care, and the exposure of sensitive records would ripple through the institution and beyond, eroding confidence in the very systems meant to enhance care delivery. In a sector already strained by resource limitations and increasing reliance on digital tools, the unchecked vulnerabilities of genAI create a perfect storm for exploitation. To ignore these risks is to invite a crisis that healthcare can ill afford.

The vulnerabilities exposed in these cases, while significant, are far from insurmountable. Healthcare institutions have an opportunity, and indeed an obligation, to adopt a multifaceted approach that safeguards genAI tools while preserving their transformative potential. The path forward demands a synthesis of technical rigor, policy innovation, and sector-wide advocacy. Robust security protocols form the foundational basis of protection against exploitation. Generative AI tools must be fortified with input validation, encryption, and MFA to shield against unauthorized access [75]. Continuous monitoring and real-time threat detection are not mere enhancements but essential defenses, enabling institutions to neutralize breaches before they escalate. The case of Vanna AI's prompt injection flaw underscores the urgency of such measures, as healthcare cannot afford a compromise in diagnostic accuracy or patient data security [55].

Policy and governance are equally critical. Clear guidelines must govern the use of genAI tools, supported by comprehensive training for clinicians and administrators on the risks of data mishandling [76, 77, 78]. The Samsung case illustrates how inadvertent errors, compounded by inadequate oversight, can lead to significant breaches [50]. Vendor agreements must enforce strict data retention and usage policies, ensuring that sensitive patient information is neither stored unnecessarily nor misused [79].

Incident response preparedness is another imperative. Tailored response plans that anticipate the unique challenges of genAI breaches can mitigate downtime and operational disruptions. Simulations and scenario-based training should be routine, ensuring that staff can respond effectively to crises [5]. For example, institutions reliant on AI-powered EHR systems must be prepared to maintain continuity of care during outages, as evidenced by the lessons of the ChatGPT breach. Technology-

specific safeguards offer a pragmatic solution to many vulnerabilities [80]. Healthcare-specific genAI tools, designed with built-in safeguards against misuse and cyberattacks, provide a safer alternative to generic systems. In-house solutions allow for greater control over data handling and reduce exposure to third-party risks, as the deepfake scam highlights the dangers of unmonitored external systems.

Finally, advocacy and regulation are indispensable. The healthcare sector must champion robust legal frameworks that address data usage, accountability, and liability in AI-driven breaches [81]. Effective regulation deters malicious actors while fostering a culture of responsibility [82]. As healthcare increasingly integrates genAI, clear standards are essential to ensure that innovation proceeds hand in hand with security [2]. These measures are not merely defensive but a proactive blueprint for the future. In navigating the complexities of genAI, healthcare must strike a delicate balance: harnessing its potential while safeguarding the patients and systems it serves. Only by addressing these vulnerabilities head-on can healthcare institutions build the trust and resilience required to thrive in an AI-driven era [83].

While this study provides a critical foundation for understanding the security risks posed by genAI in healthcare, its conclusions must be tempered by certain limitations. Foremost among these is the reliance on hypothetical scenarios extrapolated from real-world incidents outside the healthcare domain. While these analogs offer valuable perspectives, they inevitably fall short of capturing the intricate nuances of healthcare workflows, regulatory frameworks, and system complexities. Such divergences may constrain the direct applicability of the findings to clinical settings. The rapid evolution of genAI further complicates this analysis. As new models emerge and safeguards are implemented, vulnerabilities identified in this study may be mitigated, or entirely new risks may arise. This fluid landscape underscores the need for continual reassessment to ensure the enduring relevance of the insights presented here. Moreover, the focus on specific cases—such as ChatGPT data breaches and deepfake scams—while illuminating, cannot comprehensively address the diversity of genAI tools or their varied applications across healthcare domains. The study's reliance on the CIA framework, though robust, presents another limitation. This framework, by design, emphasizes technical dimensions, potentially overlooking broader issues such as ethical dilemmas, societal impacts, and the psychological toll of breaches on patients and providers. Finally, the absence of quantitative data on the prevalence and consequences of genAI breaches in healthcare limits the ability to gauge their true magnitude, necessitating further empirical research. Despite these constraints, this study establishes an essential groundwork for future exploration, offering a structured approach to understanding and mitigating the vulnerabilities inherent in genAI. As healthcare's reliance on these technologies deepens, ongoing investigation and adaptation will be critical to ensuring their safe and effective integration.

5. CONCLUSIONS

Generative AI stands at the crossroads of promise and peril for healthcare, embodying a profound duality: the capacity to revolutionize care delivery and operational efficiency, tempered by the significant risks it introduces. The case studies examined in this study illuminate the intricate vulnerabilities—confidentiality, integrity, and availability—that lie at the heart of these tools. From the exposure of sensitive data to the insidious dangers of deepfake manipulations, the stakes for healthcare institutions are unmatched in their complexity and consequence. Yet, these risks, though formidable, are not insurmountable. The lessons drawn from these incidents offer a roadmap for navigating the challenges of genAI adoption. By implementing rigorous safeguards, fostering cross-sector collaboration, and committing to proactive vigilance, healthcare can harness the transformative potential of these technologies while preserving the sanctity of patient safety and institutional trust. The path forward is neither simple nor static. It demands an unwavering dedication to ethical innovation, informed by the dual imperatives of embracing progress and guarding against its misuse. Generative AI offers a powerful tool to advance healthcare, but its promise can only be realized through deliberate action to mitigate its risks. In striking this balance, the industry not only safeguards the patients and systems it serves but also sets a standard for the responsible integration of transformative technologies across all domains.

Abbreviations

AI	Artificial intelligence
CDSS	Clinical decision support systems
CIA	Confidentiality, Integrity, and Availability
genAI	Generative artificial intelligence
HER	Electronic health records
HIPAA	Health Insurance Portability and Accountability Act
HIV	Human immunodeficiency virus

LLM Large language model
MFA Multifactor authentication
PII Personally identifiable information
RCE Remote code execution

Conflicts Of Interest

The authors declare no conflicts of interest.

Funding

The authors clearly indicate that the research was conducted without any funding from external sources.

Acknowledgment

The first draft of the manuscript was written by the author, Malik Sallam. No external writing assistance was provided. OpenAI's ChatGPT-4o model was used to refine the language and structure of the manuscript. All final edits and intellectual contributions were made solely by the authors, and no external funding or payment was involved in the writing process.

References

[1] P. Zhang and M. N. Kamel Boulos, "Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges," *Future Internet*, vol. 15, no. 9, p. 286, 2023, doi: 10.3390/fi15090286.

[2] D. Yim, J. Khuntia, V. Parameswaran, and A. Meyers, "Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review," (in eng), *JMIR Med Inform*, vol. 12, p. e52073, Mar 20 2024, doi: 10.2196/52073.

[3] R. Xu and Z. Wang, "Generative artificial intelligence in healthcare from the perspective of digital media: Applications, opportunities and challenges," (in eng), *Heliyon*, vol. 10, no. 12, p. e32364, Jun 30 2024, doi: 10.1016/j.heliyon.2024.e32364.

[4] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," (in eng), *Healthcare (Basel)*, vol. 11, no. 6, p. 887, Mar 19 2023, doi: 10.3390/healthcare11060887.

[5] S. Reddy, "Generative AI in healthcare: an implementation science informed translational path on application, integration and governance," *Implementation Science*, vol. 19, no. 1, p. 27, 2024/03/15 2024, doi: 10.1186/s13012-024-01357-9.

[6] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," (in eng), *Future Healthc J*, vol. 8, no. 2, pp. e188-e194, Jul 2021, doi: 10.7861/fhj.2021-0095.

[7] A.-H. Al-Mistarehi, M. Mijwil, Y. Filali, B. Mariem, A. L. I. Guma, and M. Abotaleb, "Artificial Intelligence Solutions for Health 4.0: Overcoming Challenges and Surveying Applications," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. 2023, pp. 15-20, 03/10 2023, doi: 10.58496/MJAIH/2023/003.

[8] M. Sallam, "Bibliometric top ten healthcare-related ChatGPT publications in the first ChatGPT anniversary," (in eng), *Narra J*, vol. 4, no. 2, p. e917, Aug 2024, doi: 10.52225/narra.v4i2.917.

[9] D. Leslie and X.-L. Meng, "Future Shock: Grappling With the Generative AI Revolution," *Harvard Data Science Review*, no. Special Issue 5, 2024, doi: 10.1162/99608f92.fad6d25c.

[10] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "Generative artificial intelligence: a systematic review and applications," *Multimedia Tools and Applications*, 2024/08/14 2024, doi: 10.1007/s11042-024-20016-1.

[11] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," (in eng), *Future Healthc J*, vol. 6, no. 2, pp. 94-98, Jun 2019, doi: 10.7861/futurehosp.6-2-94.

[12] S. A. Alowais *et al.*, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC Medical Education*, vol. 23, no. 1, p. 689, 2023/09/22 2023, doi: 10.1186/s12909-023-04698-z.

[13] K. Moulaei, A. Yadegari, M. Baharestani, S. Farzanbakhsh, B. Sabet, and M. R. Afrash, "Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications," *International Journal of Medical Informatics*, 05/08 2024, doi: 10.1016/j.ijmedinf.2024.105474.

[14] E. Urquhart *et al.*, "A pilot feasibility study comparing large language models in extracting key information from ICU patient text records from an Irish population," *Intensive Care Medicine Experimental*, vol. 12, no. 1, p. 71, 2024/08/16 2024, doi: 10.1186/s40635-024-00656-1.

[15] J. W. Ayers *et al.*, "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," (in eng), *JAMA Intern Med*, vol. 183, no. 6, pp. 589-596, Jun 1 2023, doi: 10.1001/jamainternmed.2023.1838.

[16] C. Lee, K. A. Vogt, and S. Kumar, "Prospects for AI clinical summarization to reduce the burden of patient chart review," (in eng), *Front Digit Health*, vol. 6, p. 1475092, 2024, doi: 10.3389/fdgth.2024.1475092.

[17] J. Zaretsky *et al.*, "Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format," (in eng), *JAMA Netw Open*, vol. 7, no. 3, p. e240357, Mar 4 2024, doi: 10.1001/jamanetworkopen.2024.0357.

[18] G. Sánchez-Rosenberg *et al.*, "ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases," (in eng), *Acta Orthop*, vol. 95, pp. 152-156, Mar 21 2024, doi: 10.2340/17453674.2024.40182.

[19] M. Mijwil, A. Mohammad, and A. Ahmed Hussein, "ChatGPT: Exploring the Role of Cybersecurity in the Protection of Medical Information," *Mesopotamian Journal of CyberSecurity*, vol. 2023, pp. 18-21, 02/01 2023, doi: 10.58496/MJCS/2023/004.

[20] A. L. I. Guma and M. Mijwil, "Cybersecurity for Sustainable Smart Healthcare: State of the Art, Taxonomy, Mechanisms, and Essential Roles," *Mesopotamian Journal of CyberSecurity*, vol. 4, pp. 20–62, 05/23 2024, doi: 10.58496/MJCS/2024/006.

[21] E. Ferrara, "GenAI against humanity: nefarious applications of generative artificial intelligence and large language models," *Journal of Computational Social Science*, vol. 7, no. 1, pp. 549-569, 2024/04/01 2024, doi: 10.1007/s42001-024-00250-1.

[22] A. Bandi, P. V. Adapa, and Y. E. Kuchi, "The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges," *Future Internet*, vol. 15, no. 8, p. 260, 2023, doi: 10.3390/fi15080260.

[23] Swivel Secure. "9 reasons why healthcare is the biggest target for cyberattacks." https://swivelsecure.com/solutions/healthcare/healthcare-is-the-biggest-target-for-cyberattacks/ (accessed 10 December 2024, 2024).

[24] P. Ewoh and T. Vartiainen, "Vulnerability to Cyberattacks and Sociotechnical Solutions for Health Care Systems: Systematic Review," (in eng), *J Med Internet Res*, vol. 26, p. e46904, May 31 2024, doi: 10.2196/46904.

[25] M. A. Arain, R. Tarraf, and A. Ahmad, "Assessing staff awareness and effectiveness of educational training on IT security and privacy in a large healthcare organization," (in eng), *J Multidiscip Healthc*, vol. 12, pp. 73-81, 2019, doi: 10.2147/jmdh.S183275.

[26] H. T. Neprash *et al.*, "Trends in Ransomware Attacks on US Hospitals, Clinics, and Other Health Care Delivery Organizations, 2016-2021," (in eng), *JAMA Health Forum*, vol. 3, no. 12, p. e224873, Dec 2 2022, doi: 10.1001/jamahealthforum.2022.4873.

[27] M. S. Jalali and J. P. Kaiser, "Cybersecurity in Hospitals: A Systematic, Organizational Perspective," (in eng), *J Med Internet Res*, vol. 20, no. 5, p. e10059, May 28 2018, doi: 10.2196/10059.

[28] A. H. Seh *et al.*, "Healthcare Data Breaches: Insights and Implications," (in eng), *Healthcare (Basel)*, vol. 8, no. 2, p. 133, May 13 2020, doi: 10.3390/healthcare8020133.

[29] J. Tully, J. Selzer, J. Phillips, P. O'Connor, and C. Dameff, "Healthcare Challenges in the Era of Cybersecurity," *Health Security*, vol. 18, pp. 228-231, 06/01 2020, doi: 10.1089/hs.2019.0123.

[30] A. Das, A. Tariq, F. Batalini, B. Dhara, and I. Banerjee, "Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer," (in eng), *medRxiv*, Mar 21 2024, doi: 10.1101/2024.03.20.24304627.

[31] Z. L. Teo, C. W. N. Quek, J. L. Y. Wong, and D. S. W. Ting, "Cybersecurity in the generative artificial intelligence era," *Asia-Pacific Journal of Ophthalmology*, vol. 13, no. 4, p. 100091, 2024/07/01/ 2024, doi: 10.1016/j.apjo.2024.100091.

[32] L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri, and J. Hernández-Orallo, "Larger and more instructable language models become less reliable," *Nature*, vol. 634, no. 8032, pp. 61-68, 2024/10/01 2024, doi: 10.1038/s41586-024-07930-y.

[33] M. Mijwil, M. Aljanabi, and Chatgpt, "Towards Artificial Intelligence-Based Cybersecurity: The Practices and ChatGPT Generated Ways to Combat Cybercrime," *Iraqi Journal for Computer Science and Mathematics*, vol. 4, pp. 65-70, 01/19 2023, doi: 10.52866/ijcsm.2023.01.01.0019.

[34] OpenAI. "March 20 ChatGPT outage: Here's what happened." https://openai.com/index/march-20-chatgpt-outage/ (accessed 10 December 2024, 2024).

[35] L. Wilkinson and Industry Dive. "Samsung employees leaked corporate data in ChatGPT: report." Industry Dive is an Informa TechTarget business. https://www.cybersecuritydive.com/news/Samsung-Electronics-ChatGPT-leak-data-privacy/647219/ (accessed 10 December 2024, 2024).

[36] P. Singh. "Samsung employees accidentally leaked company secrets via ChatGPT: Here's what happened." BUSINESS TODAY. https://www.businesstoday.in/technology/news/story/samsung-employees-accidentally-

leaked-company-secrets-via-chatgpt-heres-what-happened-376375-2023-04-06 (accessed 10 December 2024, 2024).
[37] K. Huang, J. Huang, and D. Catteddu, "GenAI Data Security," in *Generative AI Security: Theories and Practices*, K. Huang, Y. Wang, B. Goertzel, Y. Li, S. Wright, and J. Ponnapalli Eds. Cham: Springer Nature Switzerland, 2024, pp. 133-162.
[38] R. N. Lebow, "Nuclear Crisis Management: A Dangerous Illusion," *Bulletin of Peace Proposals*, vol. 17, no. 2, pp. 107-112, 1986. [Online]. Available: http://www.jstor.org/stable/44481231.
[39] V. Turusov, V. Rakitsky, and L. Tomatis, "Dichlorodiphenyltrichloroethane (DDT): ubiquity, persistence, and risks," (in eng), *Environ Health Perspect*, vol. 110, no. 2, pp. 125-8, Feb 2002, doi: 10.1289/ehp.02110125.
[40] E. A. Emmett, "Asbestos in High-Risk Communities: Public Health Implications," (in eng), *Int J Environ Res Public Health*, vol. 18, no. 4, p. 1579, Feb 7 2021, doi: 10.3390/ijerph18041579.
[41] G. AlMudahi, L. Alswayeh, S. Alansary, and R. Latif, *Social Media Privacy Issues, Threats, and Risks*. 2022, pp. 155-159.
[42] H. Taherdoost, "Security and Internet of Things: Benefits, Challenges, and Future Perspectives," *Electronics*, vol. 12, no. 8, p. 1901, 2023, doi: 10.3390/electronics12081901.
[43] T. Hresko Pearl, "Fast & Furious: The Misregulation of Driverless Cars," *NYU Annual Survey of American Law*, vol. 73, no. 24, 2016, doi: 10.2139/ssrn.2819473.
[44] C. Brokowski and M. Adli, "CRISPR Ethics: Moral Considerations for Applications of a Powerful Tool," (in eng), *J Mol Biol*, vol. 431, no. 1, pp. 88-101, Jan 4 2019, doi: 10.1016/j.jmb.2018.05.044.
[45] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, p. 101896, 2023/11/01/ 2023, doi: 10.1016/j.inffus.2023.101896.
[46] M. Aminzade, "Confidentiality, integrity and availability – finding a balanced IT framework," *Network Security*, vol. 2018, no. 5, pp. 9-11, 2018/05/01/ 2018, doi: 10.1016/S1353-4858(18)30043-6.
[47] N. Chitadze, "Basic Principles of Information and Cyber Security," in *Analyzing New Forms of Social Disorders in Modern Virtual Environments*, M. Boskovic, G. Misev, and N. Putnik Eds. Hershey, PA, USA: IGI Global, 2023, pp. 193-223.
[48] J. Morales and CNN.com. "OpenAI Hack: Examining ChatGPT Security Vulnerabilities." https://www.ccn.com/news/technology/openai-hack-chat-gpt-security-vulnerabilities/ (accessed 11 December 2024, 2024).
[49] S. Poremba and Security Intelligence. "ChatGPT confirms data breach, raising security concerns." https://securityintelligence.com/articles/chatgpt-confirms-data-breach/ (accessed 10 December 2024, 2024).
[50] M. Gurman and Bloomberg.com. "Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak." https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak?embedded-checkout=true (accessed 10 December 2024, 2024).
[51] S. Ray and Forbes.com. "Samsung Bans ChatGPT Among Employees After Sensitive Code Leak." https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/ (accessed 11 December 2024, 2024).
[52] VentureBeat. "A Chevy for $1? Car dealer chatbots show perils of AI for customer service." https://venturebeat.com/ai/a-chevy-for-1-car-dealer-chatbots-show-perils-of-ai-for-customer-service/ (accessed 11 December 2024, 2024).
[53] N. Ancell and Cybernews. "Chevrolet dealership duped by hacker into selling $70K car at criminally low price." https://cybernews.com/ai-news/chevrolet-dealership-chatbot-hack/ (accessed 10 December 2024, 2024).
[54] The Hacker News (THN). "Prompt Injection Flaw in Vanna AI Exposes Databases to RCE Attacks." https://thehackernews.com/2024/06/prompt-injection-flaw-in-vanna-ai.html (accessed 11 December 2024, 2024).
[55] N. Nehorai. "When Prompts Go Rogue: Analyzing a Prompt Injection Code Execution in Vanna.AI." https://jfrog.com/blog/prompt-injection-attack-code-execution-in-vanna-ai-cve-2024-5565/ (accessed 10 December 2024, 2024).
[56] K. Magramo and CNN. "British engineering giant Arup revealed as $25 million deepfake scam victim." https://edition.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/index.html (accessed 11 December 2024, 2024).
[57] D. Milmo and The Guardian. "UK engineering firm Arup falls victim to £20m deepfake scam." https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video (accessed 11 December 2024, 2024).

[58] A. Ho *et al.*, "Algorithmic progress in language models," *arXiv preprint arXiv:2403.05812*, 2024, doi: 10.48550/arXiv.2403.05812.
[59] I. Jada and T. O. Mayayise, "The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review," *Data and Information Management*, vol. 8, no. 2, p. 100063, 2024/06/01/ 2024, doi: 10.1016/j.dim.2023.100063.
[60] Y. Chen and P. Esmaeilzadeh, "Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges," (in eng), *J Med Internet Res*, vol. 26, p. e53008, Mar 8 2024, doi: 10.2196/53008.
[61] J. Li, "Security Implications of AI Chatbots in Health Care," (in eng), *J Med Internet Res*, vol. 25, p. e47551, Nov 28 2023, doi: 10.2196/47551.
[62] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," *BMC Medical Ethics*, vol. 22, no. 1, p. 122, 2021/09/15 2021, doi: 10.1186/s12910-021-00687-3.
[63] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," *Information Fusion*, vol. 97, p. 101804, 2023/09/01/ 2023, doi: 10.1016/j.inffus.2023.101804.
[64] A. Abernethy *et al.*, "The Promise of Digital Health: Then, Now, and the Future," (in eng), *NAM Perspect*, vol. 2022, 2022, doi: 10.31478/202206e.
[65] M. Zarour *et al.*, "Ensuring data integrity of healthcare information in the era of digital health," (in eng), *Healthc Technol Lett*, vol. 8, no. 3, pp. 66-77, Jun 2021, doi: 10.1049/htl2.12008.
[66] M. Franklin, P. M. Tomei, and R. Gorman, "Strengthening the EU AI Act: Defining Key Terms on AI Manipulation," *arXiv preprint arXiv:2308.16364*, 2023, doi: 10.48550/arXiv.2308.16364.
[67] A. J. Rivera-Rodriguez and B. T. Karsh, "Interruptions and distractions in healthcare: review and reappraisal," (in eng), *Qual Saf Health Care*, vol. 19, no. 4, pp. 304-12, Aug 2010, doi: 10.1136/qshc.2009.033282.
[68] T. P. Quinn, M. Senadeera, S. Jacobs, S. Coghlan, and V. Le, "Trust and medical AI: the challenges we face and the expertise needed to overcome them," (in eng), *J Am Med Inform Assoc*, vol. 28, no. 4, pp. 890-894, Mar 18 2021, doi: 10.1093/jamia/ocaa268.
[69] M. Muthuppalaniappan and K. Stevenson, "Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health," (in eng), *Int J Qual Health Care*, vol. 33, no. 1, Feb 20 2021, doi: 10.1093/intqhc/mzaa117.
[70] M. I. Mashinchi, T. Acton, and P. M. Datta, "When healthcare becomes sick: Recovering from ransomware," *Journal of Information Technology Teaching Cases*, p. 20438869241279443, 2024, doi: 10.1177/20438869241279443.
[71] G. Moore, Z. Khurshid, T. McDonnell, L. Rogers, and O. Healy, "A resilient workforce: patient safety and the workforce response to a cyber-attack on the ICT systems of the national health service in Ireland," *BMC Health Services Research*, vol. 23, no. 1, p. 1112, 2023/10/17 2023, doi: 10.1186/s12913-023-10076-8.
[72] J. Zarocostas, "Health under cyberattack," *The Lancet*, vol. 398, no. 10303, pp. 829-830, 2021, doi: 10.1016/S0140-6736(21)01968-1.
[73] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy," *IEEE Access*, vol. 11, pp. 80218-80245, 2023, doi: 10.1109/ACCESS.2023.3300381.
[74] O. Navarro Martínez, D. Fernández-García, N. Cuartero Monteagudo, and O. Forero-Rincón, "Possible Health Benefits and Risks of DeepFake Videos: A Qualitative Study in Nursing Students," *Nursing Reports*, vol. 14, no. 4, pp. 2746-2757, 2024, doi: 10.3390/nursrep14040203.
[75] H. Hayagreevan and S. Khamaru, "Security of and by Generative AI platforms," *arXiv preprint arXiv:2410.13899*, 2024, doi: 10.48550/arXiv.2410.13899.
[76] B. L. Moorhouse, M. A. Yeo, and Y. Wan, "Generative AI tools and assessment: Guidelines of the world's top-ranking universities," *Computers and Education Open*, vol. 5, p. 100151, 2023/12/15/ 2023, doi: 10.1016/j.caeo.2023.100151.
[77] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *npj Digital Medicine*, vol. 6, no. 1, p. 120, 2023/07/06 2023, doi: 10.1038/s41746-023-00873-0.
[78] S. T. Argaw, N.-E. Bempong, B. Eshaya-Chauvin, and A. Flahault, "The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 10, 2019/01/11 2019, doi: 10.1186/s12911-018-0724-5.
[79] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, and J. Qadir, "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Computers in Biology and Medicine*, vol. 158, p. 106848, 2023/05/01/ 2023, doi: 10.1016/j.compbiomed.2023.106848.
[80] R. Rodrigues, "Legal and human rights issues of AI: Gaps, challenges and vulnerabilities," *Journal of Responsible Technology*, vol. 4, p. 100005, 2020/12/01/ 2020, doi: 10.1016/j.jrt.2020.100005.

[81] D. Ueda *et al.*, "Fairness of artificial intelligence in healthcare: review and recommendations," (in eng), *Jpn J Radiol*, vol. 42, no. 1, pp. 3-15, Jan 2024, doi: 10.1007/s11604-023-01474-3.

[82] M. Balas, D. T. Wong, and S. A. Arshinoff, "Artificial intelligence, adversarial attacks, and ocular warfare," *AJO International*, vol. 1, no. 3, p. 100062, 2024/10/03/ 2024, doi: 10.1016/j.ajoint.2024.100062.

[83] A. Garcia-Perez, J. G. Cegarra-Navarro, M. P. Sallos, E. Martinez-Caro, and A. Chinnaswamy, "Resilience in healthcare systems: Cyber security and digital transformation," *Technovation*, vol. 121, p. 102583, 2023/03/01/ 2023, doi: 10.1016/j.technovation.2022.102583.