Cyber Security Issues and Challenges Related to Generative AI and ChatGPT

1st Rajesh Pasupuleti,
Assistant Scientist,
Frost Institute for Data Science and Computing (IDSC),
University of Miami,
Coral Gables, FL, USA

2nd Dr Ravi Vadapalli, Ph.D
Faculty Director,
IDSC Advanced Computing Systems,
Research Associate Professor,
Electrical and Computer Engineering (ECE),
University of Miami,
Coral Gables, FL,USA

3rd Dr Christopher Mader,
Sr. Director,
Systems and Data Engineering,
Frost Institute for Data Science and Computing (IDSC),
University of Miami,
Coral Gables, FL, USA

***Abstract*—In recent years, Generative Artificial Intelligence (AI) and ChatGPT (Generative Pre-trained Transformer) models that are capable of generating realistic human-mimicked languages have gained progressive popularity. With the evolution of technology, there has been a significant increase in the availability and usage of artificial intelligence tools, such as ChatGPT and Generative AI, that will assist in shaping the future. However, this increasing popularity poses a potential risk if used inappropriately. Threats from AI pose special challenges for government, the private sector, and national security. In this paper, we address some of key concerns of significant cyber security issues and challenges related to Generative AI and ChatGPT. With careful consideration to application usage, organizations can implement appropriate security measures to mitigate these risks. We also incorporate recommendations about ChatGPT usage and its impact on society. It is important that researchers, developers, and policymakers (CIOs, CSOs) work together to mitigate these risks and to ensure that these models are used in a responsible and ethical manner.**

Keywords—Cybersecurity Issues, Challenges, Generative AI, ChatGPT, LLM Models, Risks

I. INTRODUCTION

As technology continues to rapidly evolve, global regulations undergo changes, and public confidence in organizations is challenged, a more comprehensive approach to strategic planning is necessary. Generative AI has improved decision making and enhanced personalization and the contextual information retrieval of results by remembering user queries, but it also poses various risks to society. Perhaps any new technological innovation causes seventy percent excitement and thirty percent nervousness due to the exposure to threats in real time.

If AI machine intelligence controls human beings and starts thinking and taking corresponding actions, it would be a profound risk to society. Algorithms play an instrumental role in both machine learning and generative AI models, optimizing the accuracy of outputs while simultaneously making decisions or suggestions based on input data. This allows machines to learn from data inputs and to make reliable decisions or recommendations based on these inputs .

We have seen already cases of the revolution of generative AI and its impacts on society addressed by the President of the U.S. at the White House, along with the CEOs of top companies like Microsoft, Google, Meta, and OpenAI, as they considered how to control artificial intelligence technology and how to create secure and efficient policy standards to safeguard the utilization of technology from malicious actors and attacks [1].

AI technologies are advancing quickly, typically outpacing generative AI and ChatGPT within a five-month span of reaching audiences. Assessments of the technologies, consequences of every business organizations requires expertise control board concentrated in the private sector along with policies and organizational reforms within government [2].

Cyberhaven found 11% of its employees had posted company data on ChatGPT, and three engineers at Samsung recently shared sensitive corporate information with ChatGPT in order to find errors in the semiconductor code and to optimize Samsung's equipment code [3].

"Generative AI represents a radical shift in how companies from all industries will analyze data, automate processes, and equip sales, service, marketing, and commerce professionals with the tools to enhance customer relationships; nevertheless, it comes with its own unique risks and challenges." said Clara Shih, CEO of Service Cloud Salesforce[4].

The global cyber security market is expected to witness substantial growth in the coming years. Statista data suggests that global cybersecurity market revenue is projected to skyrocket to $262.3 billion by 2027, an increase of 67% from this year's total of $156.35 billion.[5][6].

A survey among Thales consumers revealed that the potential security risks stemming from generative AI tools provoked worries in 75% of respondents. It included: the risk of exposing sensitive information, the potential to create malicious code, divulging personal information, spreading fake news, creating more convincing phishing emails, and the threat of disinformation. In 2023, phishing attacks were weaponized and spreading its wings across commonly used apps and business communication services[7]. "In the future, it might spread across communication channels in a much stealthier way beyond email and messages and into geopolitically motivated cyberattacks" said Jaspreet Singh of Trellix, also based in California [8].

Zoombombing is the hijacking of video conference calls by hackers in order to gain email addresses and identities. The

architectural flow and functionality of products has been observed as the use of business collaboration apps grow as threat vectors.

ChatGPT also experienced a significant bug that leaked user conversation histories. Widespread apprehension regarding data privacy led Italy to become the first Western nation to completely prohibit ChatGPT while they investigate any possible data privacy infractions [9][10].

ChatGPT sources the public web to generate data. It's easy to trace the exact source employed in model creation, also known as a "training data extraction attack". Large, billion-parameter language models that have been trained on public internet or private datasets leads to data extraction attacks[11]. The repercussions of the generation of data includes the exposure of sensitive, personally identifiable information such as phone numbers, mails, address, etc.

Proponents of ChatGPT and Generative AI could be outweigh potential risks through the implementation of inappropriate security protocols and controls employed by hackers . The author "Biswas" also discussed various kinds of information leakage risks in chatbot applications including data interception, data leakage, unauthorized access, financial loss, organizations reputational damage and legal liabilities [12][13].

AI TRISM (Trust, Risk, and Security Management) supports the urgent need for developing a new AI model that will ensure trustworthiness, governance, reliability, fairness, efficacy, robustness, and data protection. It incorporates methods for describing AI-generated results, rapidly deploying new innovative models, and actively managing artificial intelligence security controls for privacy and Ethics issues [14].

The accelerated rise of ChatGPT vibrates across multiple industries including law. In recent news from CNN, according to a New York court one attorney was accused of citing approximately six fake cases obtained using ChatGPT to support and strengthen his legal arguments. Opposing counsel realized the fabrications and challenged the citations. The lawyer was unaware of the misinformation of content produced by ChatGPT (a.k.a. "hallucinations")[15].

ChatGPT, generative AI, even coding Github Copilot, or Codex heralds the beginning of a new era in human-machine collaborations. The next section describes cyber security issues and challenges related to Generative AI and ChatGPT risks that fell under five substantial areas of modules of Data Breaches, Data Privacy, people, Malware, and Phishing Attacks.

II. Cyber Security Substantial Areas of Risks with Generative AI and Chat GPT

ChatGPT, a chatbot driven by artificial intelligence, was created to mimic human conversation and to collect information to train model, improve, and enhance the platform's performance.

Generative AI, such as chatbots powered by models like GPT, brings about several security issues and challenges. Here are some of the key concerns:

Misinformation and Manipulation It is important to note that Generative AI can be used to spread misinformation and manipulation. Malicious actors can exploit chatbots to disseminate propaganda, create fake news, fake social media posts, and even deepfake videos to deceive individuals. This poses a significant challenge to maintaining accurate and reliable information. One of the most significant concerns with Generative AI and ChatGPT models is the potential for malicious use.

Hallucinations That Threaten Performance Artificial intelligence technologies are exposed to use in disinformation campaigns. Deepfake images and videos represent an obvious threat. Generative text and voice cloning also merit concern.

Privacy and Data Security Furthermore, Generative AI and ChatGPT require large amounts of data for effective training, data that could potentially contain sensitive personal information such as email addresses, phone numbers, or credit card details. Therefore, security measures must be taken in order to protect user information from potential misuse. Protecting this data from unauthorized access or breaches is crucial. Inadequate security measures may lead to unauthorized data leaks or privacy violations.

Phishing and Social Engineering Attacks Generative AI and ChatGPT can be manipulated to engage in phishing mail attacks or social engineering attempts. Malicious actors may exploit the trust placed in chatbots and trick users into revealing sensitive information, such as passwords or financial details.

Bias and Discrimination Generative AI models such as ChatGPT are capable of inheriting bias that is present in the training data. This can lead to unintentional and potentially unfair outcomes, with certain individuals or groups being unfairly treated or marginalized when the underlying data contains discriminatory content.

Offensive or Inappropriate Content Without proper content filtering and moderation mechanisms, chatbots powered by generative AI can produce offensive or inappropriate responses. This can result in negative experiences for users and damage the reputation of the organizations deploying such chatbots.

Deepfakes and Identity Theft In addition, Generative AI is capable of producing convincing deep fakes: videos, images, and other works of art that are artificially generated but appear to be authentic. Chatbots might contribute to this by generating text or images that impersonate individuals, leading to identity theft or fraudulent activities.

Adversarial Attacks Adversarial attacks involve intentionally manipulating inputs to trick the AI system into producing incorrect or unexpected outputs. Chatbots powered by generative AI models can be vulnerable to such attacks, which could result in misleading or harmful responses.

HIPAA Compliance ChatGPT is not HIPAA compliant (Health Insurance Portability and Accountability Act of 1996), and University, medical health data representatives, and **PHI** (Protected Health Information) members do not permit its use with any sensitive patient data and are developing a process/policy to ensure that appropriate use in health care and universities is necessary.

Liability OpenAI, the company behind ChatGPT, also warns against feeding confidential and sensitive data into the platform that could constitute a public disclosure and lead to

a loss and liability in terms of an individual or company. This also leads to repercussions of events related to confidentiality of employment.

Cyber Security Issues Generative AI and ChatGPT are beneficial for developers in an ethical sense, but they pose security risks for organizations as malicious code and phishing emails can be created with greater efficiency than ever before, and the potential AI generation of code may also pose serious concerns to securing Kubernetes and cloud environments. Research has already demonstrated how to exploit a buffer overflow, congestion control, and code injection into a common executable to delay the performance of applications.

Debugging Generative AI platforms can be adept at debugging and finding programming errors and software code, it can also leads to an attacker could make use of this capability to find security vulnerabilities.

Model Interpretation ChatGPT and Generative AI will learn from model training and remember the previous history of query results. Supposing hackers train the model wrongly, the further consequences of output results will be misleads on the original context.

Job Displacement The utilization of Generative AI and ChatGPT technology may lead to job displacement in areas such as technical writing and resume building, which could have a profound effect on workers and their communities if no other job opportunities are available.

Accountability Given the obscurity of Generative AI systems, it is difficult to assign culpability for mistakes or errors when something goes wrong. This raises issues about accountability and the potential to hold people or organizations responsible for the outcomes of decisions created by AI.

Safety Utilizing generative AI in machine-critical control systems or industrial control embedded systems such as aircraft design or air traffic control raises significant worries regarding system safety of high priority. Errors or prejudices present in an AI system could prove fatal, leading to accidents or other safety incidents.

Business Impact ChatGPT is still in its infancy. Organizations are creating security policies and guidelines to provide secure business services, and some companies restrict the use of ChatGPT altogether at the workplace. Amazon, JP Morgan, Walmart have given warnings to their employees about the use of AI services and about keeping sensitive information in ChatGPT.

Sensitive Information There are inherent risks in feeding sensitive information to a large language model. The data could be retrieved by a different person or group of people for later use.

Different Scenarios of Confidential Data Leaks on ChatGPT An executive submitted their company strategy document into ChatGPT, requested a query to make a slide deck out of it. Another scenario involved an account of a doctor feeding his patient's name, medical details, and condition, asking ChatGPT to create a letter to the patient's insurance carrier.

Prompt Injection, Jailbreaking ChatGPT and AI technology has shown that it's vulnerable to a new attack called "prompt injection," which is the process of hijacking a language model's output. It allows the hacker to get the model to say anything that they want. It can also lead to "jailbreaking," which is the process of exploiting locked-down installed software (other than what the manufacturer has made available for that model or device) and gives hackers has full access to all the features and control model.

Control Limitation Generative AI techniques can be unstable at times, leading to unpredictable behavior. For example, Generative Adversarial Networks (GANs) may generate outputs that do not meet expectations, without providing an understandable explanation from a human perspective.

Health Data Privacy Concern The use of Generative AI in healthcare applications carries the potential risk of data privacy infringement, as it involves collecting an individual's sensitive personal information.

Religious Perspective and Ethics Issues Guidelines must be established when utilizing Generative AI innovations so as to take into account religious perspectives, and potential social, ethical, and regional aspects of generation of data, such as generating religious images, transformations of human faces into opposite genders, and mapping boarder regions.

Computational Power and Data Quantity/Quality Generative AI requires significant hardware resources such as powerful GPUs and large amounts of memory. These components tend to be expensive, which creates a hurdle for many organizations from building their own solutions.

Time Consuming and Expensive Developing a Generative AI model involves massive amounts of data and computing resources. Unfortunately, this process can be both time-consuming and costly due to the need for large datasets for training models and for generating high-quality data.

Political Aspect Organizations or individuals may train ChatGPT to generate news articles that promote a particular political agenda and spread misinformation about a specific topic, making it difficult to identify fake news once it's spread around.

Intellectual Property, Plagiarism, Copyright Issues In addition, ChatGPT has sourced the public web to generate data, which may include intellectual property, plagiarism, copyright issues, client data, business financial information extraction, or specific data privacy regulations. ChatGPT poses potential copyright issues for authors, artists, and users of the technology. It can generate text that is similar to existing works, making it possible to plagiarize content or create derivative works without permission from the original author. Additionally, its ability to generate copyrighted material could be used to unlawfully distribute such material. Furthermore, due to the complexity of AI models like ChatGPT, there is a lack of transparency regarding how it functions and generates output. All these points highlight why proper precautions must be taken when utilizing ChatGPT in order to avoid any copyright infringement.

State Law Pertaining to AI For instance, the Illinois Biometric Information Privacy Act (BIPA) mandates that companies must receive consent from individuals before collecting or utilizing their biometric data, such as fingerprints or facial recognition data. Additionally, the law requires companies to disclose how this information will be used and to uphold a data retention plan. These mandates help

promote transparency and accountability in the use of biometrics by AI systems.

These laws also bring forth the significance of transparency and responsibility when designing and deploying AI systems like ChatGPT. As AI technologies become more commonplace in our future day-to-day activities, it is indispensable that developers, users, and regulators collaborate about new innovation technologies.

III. Recommendations to Secure Usage

Organizations Site Reliability Engineers (SRE) should monitor and look out for unsanctioned utilization of ChatGPT and related technology solutions accompanied by existing cyber security controls, services, and the visualization of dashboards to catch policy violations as part of a Security Information and Event Management system solution (SIEM). In order to monitor unsanctioned utilization of ChatGPT, AI-related SRE technology can be employed for routine tasks, such as: monitoring systems for problems, alerting team members when issues are detected, and executing scripts, or taking other actions in response, all of which can be automated. AI-related SRE can also be utilized for predictive maintenance to anticipate possible faults and for root-cause analysis to pinpoint the origin of an issue. Moreover, AI-related SRE is capable of identifying potential system improvements and proposing optimization strategies that enable SREs to optimize their systems more proficiently. AI-related SRE has the capacity to dramatically improve the efficiency and productivity of SREs, allowing them to manage and maintain complex systems more effectively.

Ensure Employees are Not Leaking Business Data Large corporations including Microsoft, Amazon, and Walmart recently issued warnings to their employees regarding the use of large language model (LLM) based apps, and alerted employees (on-premises or remote workers), contractors, and partners, to not share any sensitive personal or organizational information, nor any business communication conversations with ChatGPT, including B2B services, products, integration, development, or production deployments.

Prohibit Entering Sensitive Information Prohibit feeding sensitive information into ChatGPT or any LLM. Ensure that all employees know the risks of leaking confidential business data, personal identifiable information, proprietary information, intellectual property, or trade secrets into ChatGPT and AI chatbots. This includes architectural models, financial shares data, or documentation of internal resources that are not intended to be published on the Web.

Follow AI and LLM Recommendations Reading up on guidelines can help inform a security posture. For instance, ChatGPT creator OpenAI's User Guide clearly states "We are not able to delete specific prompts from your history. Please don't share any sensitive information in your conversations."

Implementation of Policies Governing AI Services The Chief Information Officer (CIO) and Chief Security Officer (CSO) of organizations are responsible for drafting and adopting policy documents and creating awareness of the potential risks of sharing sensitive information, and other related challenges.

Digital Trustworthy Technologies: ChatGPT or Generative AI having broader goals of society by determining a balance between accountability, privacy of individuals, security, and reliability ethical and responsible use. **Strategies Enhanced Model Verification:** Propose strategies for rigorous model verification to identify and mitigate potential vulnerabilities before deployment.

Collaboration with the Cybersecurity Community: Emphasize the importance of collaboration between AI developers and cybersecurity experts to address emerging threats and vulnerabilities

Continuous Monitoring: Advocate for continuous monitoring of model behavior in real-time to detect and respond promptly to any deviations from expected norms.

Personally Identifiable Information PII Be cautious with personally identifiable information (PII). Refrain from sharing or requesting personally identifiable information such as full names, addresses, social security numbers, credit card details, or passwords. AI models don't have built-in mechanisms to secure or handle such information. Avoid relying on the model for legal, medical, or financial advice, and use human expertise when necessary.

IV. Conclusion and Future Scope

Addressing these security issues and challenges requires a multi-faceted approach. It involves robust data curation and bias mitigation techniques during the model training process, implementing strong privacy and data protection measures, employing content moderation and filtering mechanisms, and continuously monitoring and updating the AI system to detect and mitigate potential vulnerabilities. Additionally, user education and awareness are essential to help individuals identify and avoid potential risks associated with generative AI and chatbots. This manuscript can also help to create awareness for utilization with careful consideration of sensitive data and it is important that researchers, developers, and policymakers work together to mitigate these risks and ensure that these models are used in a responsible and ethical manner. Thus, CEO, CIO, and CSO leaders must carefully consider its adoption and implement new policies to address possible security violations, challenges, and enterprise risks and approaches.

References

1. Press Note on May 04, 2023. "White House Meeting with CEOs on Advancing Responsible Artificial Intelligence Innovation", https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/readout-of-white-house-meeting-with-ceos-on-advancing-responsible-artificial-intelligence-innovation/.
2. Metz, Cade, and Gregory Schmidt. "Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'." *The New York Times*, The New York Times, 29 Mar. 2023, https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.
3. February 28, 2023 cyberhaven Reprt , https://www.cyberhaven.com/blog/4-2-of-workers- have-pasted-company-data-into-chatgpt/ https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/.
4. Clara Shih, CEO of Service Cloud, Salesforce, March 6, 2023, "IT Leaders Call Generative AI a 'Game Changer' but Seek Progress on Ethics and Trust". https://www.salesforce.com/news/stories/generative-ai-research/.
5. Chris Harris | EMEA Technical Director, April 4, 2023, "The Security Challenges of Generative AI Tools: Can a Loose Prompt Sink Your Ship?" https://cpl.thalesgroup.com/blog/access-management/ai-cybersecurity-challenges.

6. Size of cybersecurity market worldwide from 2021 to 2027, Published by Statista Research Department, Mar 31, 2023, https://www.statista.com/statistics/595182/worldwide-security-as-a-service-market-size/#main-content.
7. Damian Brady, April 7, 2023, "What developers need to know about generative AI", https://github.blog/2023-04-07-what-developers-need-to-know-about-generative-ai/.
8. Jaspreet Singh, Trellix, California, "2023 Threat Predictions" geopolitically motivated cyberattacks, https://www.trellix.com/en-us/assets/docs/2023-threat-predictions-cobranded.pdf.
9. Michael Kan, March 20, 2023, "ChatGPT Users Report Seeing Other People's Conversation Histories", https://www.pcmag.com/news/chatgpt-users-report-seeing-other-peoples-conversation-histories.
10. Ryan Brown, CNBC, April 4 , 2023, "Italy became the first Western country to ban ChatGPT. Here's what other countries are doing", https://www.cnbc.com/2023/04/04/italy-has-banned-chatgpt-heres-what-other-countries-are-doing.html.
11. Nicholas Carlini, Florian Tramer, Eric Wallace, etc "Extracting Training Data from Large Language Models" **Cornell University** , Computer Science, Cryptography and Security, last revised 15 Jun 2021 , https://arxiv.org/abs/2012.07805.
12. Biswas, D. (2020, December). "Privacy Preserving Chatbot Conversations". In 2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (pp. 179-182). IEEE, https://ieeexplore.ieee.org/document/9355474.
13. Biswas, S. (2023). "Prospective Role of Chat GPT in the Military: According to ChatGPT", *Qeios*. https://www.qeios.com/read/8WYYOD
14. Groombridge, David, October 17, 2023, "Gartner Top 10 Strategic Technology Trends for 2023", https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2023.
15. Ramishah Maruf, CNN news, May 28, 2023, "Lawyer apologizes for fake court citations from ChatGPT", https://www.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html.