NFSU JOURNAL OF CYBER SECURITY & DIGITAL FORENSICS

NFSU – Journal of Cyber Security and Digital Forensics
Volume –3, Issue – 1, June 2024
E – ISSN – 2583-7559

GEN AI: A SURVEY OF SECURITY AND PRIVACY THREATS

Akhil Jain

Department of Computer Engineering
Shah and Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India
akhil.jain15998@sakec.ac.in

Vasistha Ved

Department of Computer Engineering,
Shah and Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India
vasistha.ved15891@sakec.ac.in

Anish Khadtare

Department of Computer Engineering,
Shah and Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India
anish.khadtare15636@sakec.ac.in

Jainam Chheda

Department of Information Technology,
Shah and Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India
jainam.chheda16000@sakec.ac.in

Sarthak Deshmukh

Department of Information Technology,
Shah and Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India
sarthak.deshmukh15712@sakec.ac.in

Dr. Nilakshi Jain

HoD, Cybersecurity Department,
Shah and Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India
nilakshi.jain@sakec.ac.in

***Abstract*—This paper provides a brief overview of the generation of artificial intelligence (generation AI), explaining its models, applications and operational processes. It explores the role of genetically modified artificial intelligence in cybersecurity and demonstrates its benefits and potential risks to the defender. The focus has been on emerging threats posed by AI generation. Also, a case study involving misleading advertisements illustrates the threat of deep-fakes. The study emphasizes the challenges faced by platforms in detecting advanced deep-fake content and regulatory obstacles in combating false advertising. The paper essentially emphasized the complexity of the Gen AI, its impact on security and the new serious threats and challenges it posed onto the realm of security.**

Keywords—Artificial Intelligence Generation, Generation AI, Cybersecurity, Defender, Emerging Threats, Deep-fake Threats, Detection Challenges, Regulatory Obstacles, False Advertising, Security Impact.

I. INTRODUCTION

Gen AI, short for Generative Artificial Intelligence, focuses on developing models and algorithms capable of generating new content that is identical to, or comparable to human-generated data. Large datasets are fed into these AI systems, which then use their newly acquired patterns and structures to generate new text, image, audio, and other types of data. Because of its numerous applications and revolutionary potential, Gen AI has drawn a lot of interest and traction.

Types of Gen AI models:

Generative AI refers to various models and technologies aimed at producing data or content similar to human created data, having their own unique approach to content generation. The most significant categories of Gen AI models include:

A. Generative Adversarial Networks (GANs):

GANs are comprised of two neural networks—the discriminator and the generator—that were trained using the adversarial learning theory and compete in a two-player zero-sum game. Estimating the potential distribution of reality data sample populations and producing new samples based on this distribution is the aim of GANs.[1] The discriminator must determine which synthetic data—such as text, sound, or images—are true and which are false. The generator creates this synthetic data from random noise. The objective of the generator is to generate data that is more and more realistic in order to fool discriminators, while discriminators work to become more adept at telling created data from genuine data. Thanks to this competition, GANs—which have shown great promise in image synthesis, art creation, and video production—can generate remarkably realistic content.

B. Variational Autoencoders (VAEs):

VAE learns the data distribution in a way that new, relevant data can be produced from the encoded distribution.[2] VAEs are trained to encode data into a hidden space and then decode it once more to recreate the original data. They are able to generate new samples using the learned distribution after they have mastered the probabilistic representation of the input data One way to

think of the VAE is as a hybrid encoder and decoder Bayesian network.[2]

C. Autoregressive Models:

Elements of data are generated by autonomous models, each one anticipating the next based on the context of the prior one. From the anticipated probability distribution, they take a sample. Prominent instances comprise language models such as GPT, which are adept at generating coherent and contextually appropriate text.

D. Recurrent Neural Networks (RNNs):

RNN is specialized for sequence data like natural language or time series. It predicts the subsequent sequence element depending on the previous one, which makes it excellent at generative tasks. Yet, RNNs face challenges with longsequences due to the vanishing gradient issue. In order to tackle this, more sophisticated variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) are introduced, overcoming the drawbacks of conventional RNNs.

E. Transformer-based Models:

Transformers have become more and more common for tasks involving natural language processing and generation, much like the GPT series. Leveraging attention mechanisms, they effectively capture relationships among sequence elements. Transformers are excellent at managing long sequences and work in tandem, which makes them ideal for producing texts that make sense within their context.

Working of Generative AI:

The work of Gen AI involves training a model to learn the

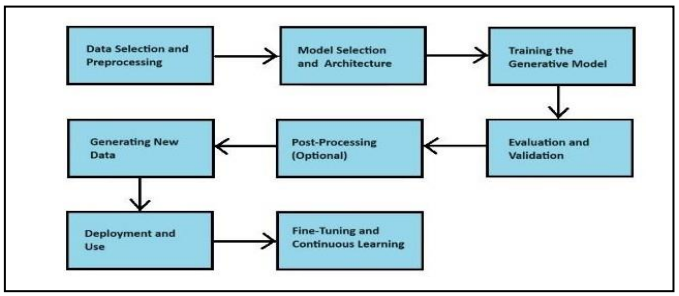

Figure 10: Flowchart of Gen AI

underlying patterns and structures within a dataset and then using the learnt knowledge to generate new, similar data. Here is a step-by-step breakdown of how Gen AI works:

Gathering and Preprocessing Data: The process begins with the collection of a dataset containing examples of the type of data that generative AI will encounter. This dataset should be reprehensible and diverse, containing a wide range of variations within the data area. The collected data are preprocessed, usually entailing steps like cleaning, normalization, and feature extraction. The objective is to get the data ready in a way that the generative model can be trained on it.[3]

Model Selection and Architecture: Depending on the particular task and the type of input, a generative model is selected. Common generative modulation types include VAEs, GANs, RNNs, Transformers and more. The architecture of the chosen generative model is defined. This includes determining the number of layers, the network structure and all additional components needed for the task.

Training the Generative Model: The generative model, trained on preprocessed data, learns statistical patterns. In adversarial training, a generator produces samples, and a discriminator distinguishes real from generated data. Iteratively adjusting model parameters minimizes a loss function, enhancing the generator's performance over time.

Generating New Data: The generative model can produce new data if it has been taught and is proficient in the distribution of unprocessed data. Random noise or noise is frequently supplied as an input to the model in order to produce new data samples. Next, this input is transformed into data samples that correspond to training data. As the model is improved and the training goes on, the quality of the generated data changes with time. [3]

Post-Processing (Optional): In some cases, the generated data may undergo post-processing steps to refine or improvise the generated output. This step depends on the specific application.

Evaluation and validation: An assessment of the generated data's quality is required. The degree to which the generated data matches the attributes of the training data can be evaluated using a variety of metrics and techniques. In order to make sure that the generated data is appropriate for its intended use and does not contain undesirable characteristics like biases or errors, validation is essential.

Use and Deployment: The generative AI model can be used for a variety of purposes after it has been trained and verified. These applications cover a wide range, including data augmentation and content generation. [3]

Fine-Tuning and Continuous Learning: In similar cases, generative AI modulators can undergo continuous learning and fine-tuning to adapt to changes in data patterns and improve their performance over time.

This paper surveys the ways in which this technology can jeopardize the aspects and virtues of security and privacy. Section III discusses the threats, and Section IV makes the concluding remarks.

I. Ease of ROLE OF GEN AI IN CYBERSECURITY

II. ROLE OF GEN AI IN CYBERSECURITY

GenAI tools like ChatGPT have a nuanced impact on cybersecurity that benefits both defenders and attackers. Defenders use them to harden systems and gain insight into comprehensivethreat intelligence. These tools improve threat intelligence based on language patterns, identifying vulnerabilities and new threats. They also help develop

security conscious human behavior and help code securely. GenAI tools play a key role in strengthening overall cybersecurity by creating secure code and test cases for validation, and provide a multifaceted approach to protecting your system from evolving cyber threats. Below, we explore the expanding roles of generative AI in cybersecurity.

A. Threat Intelligence:

GenAI boosts threat intelligence by generating concise reports on cybersecurity threats, analysing diverse sources for security professionals. Efficiently processingextensive data, it automatically detects and assesses potential threats, recommending countermeasure strategies. GenAI also analyzes security data to find trends and patterns in threat activity. [4]

B. Phishing Detection and Response:

Generative AI is employed to craft convincing phishing emails, but it can also be used to detect them. AIsystems can analyze email text and headers to identify phishing attempts by detecting inconsistencies and suspicious content.

C. Secure Code Generation and Detection:

Software availability, confidentiality, and integrity are at risk due to security flaws in code. Code review procedures are essential for spotting possible security flaws in order to fix this. By creating test cases to verify code security and generating secure code, GenAI technologies make a contribution. Additionally, LLM models support the creation of moral standards that strengthen a system's cybersecurity protections. [4]

D. Threat Scenario Simulation:

Generative AI can simulate various threat scenarios, allowing organizations to test their cybersecurity defenses. This helps in identifying weaknesses and improving incident response plans.

E. Predictive Analysis:

Utilizing machine learningalgorithms, generative AI can predict potential security breaches based on historical data and emerging trends. This proactive approach aids organizations in addressing vulnerabilities before they are exploited.

III. CYBER THREATS DUE TO GEN AI

Artificial intelligence can be misused in a variety of ways, from targeted fraud campaigns to novel forms of surveillance. A group of researchers made the decision to prioritise our top concerns when ranking the possible illegal uses of AI over the span of next 15 years. The use of Gen AI poses a number of risks to privacy and security.

A. Privacy threats:

Models are susceptible to various privacy hazards and assaults if they are not trained using privacy-preserving techniques. Because generative AI creates new data that is contextually comparable to the training set, it is crucial to make sure that sensitive information is notpresent in the training set. As AI models are trained on large databases collected from various sources that contain private information, often without the consent of the individual, there is a chance that they will inadvertently produce content that breaches an individual's private and especially the sensitive information. Because these massive models memorize significant amounts of training data, including sensitive data, which may be inadvertently disclosed or exploited maliciously by attackers, they offer privacy hazards.

B. Adversarial Attacks:

The purpose of adversarial attacks is to manipulate input data in a way that may be undetectable to humans but leads to inaccurate predictions or classifications by AI algorithms. Attackers can take advantage of weaknesses in the model's decision limits by purposefullyaltering input data. A variety of machine learning models, such as those for natural language processing and image classification, are susceptible to adversarial attacks. Small perturbations added to the input data are onetechnique that might cause misclassifications and raise security and safety issues.

C. Poisoning Data:

Data poisoning is the practice of introducing erroneous or biassed data into an artificial intelligence model's training set. The model may learn and reinforce unwanted patterns or producebiassed predictions if it is trained on corrupted data. It's imperative to guarantee the integrity andquality of training data. Attackers might try to alter the training set in order to add biases, impair the model's functionality, or even build backdoorsthat could be used in the future. In order to reduce this risk, strong procedures for data validation andcleansing are necessary.

D. Vulnerabilities in AI Models:

Like any programme, AI models can contain flaws that hackers could take advantage of. Thesevulnerabilities may result from implementation issues, code mistakes, or defects in the underlying algorithms. To find and fix possible vulnerabilities, it's critical to carry out in-depth security audits of AI models. To reduce known vulnerabilities, fixes and updates should beinstalled on a regular basis. Adversarial testing can also be used to assess how resilient the modelis against possible attacks.

E. Abnormal Reliance:

Serious problems can arise from relying too muchon AI systems without sufficient humansupervision. Unexpected events could be difficultfor automated systems to manage, necessitating human intervention to guarantee proper decision-making. Artificial intelligence (AI) systems can be a great help, but they shouldn't completely replace human judgement. In order to handle unforeseen events, understand unclear situations, and make moral decisions, human oversight is necessary. For the appropriate and efficient use

ofAI, a balance between automation and human interaction is essential.

F. Malware Generation:

There is increasing concern about these AI systems' potential to create harmful code and takeadvantage of holes in computer systems as they get more advanced. For cybersecurity experts,this threat presents a serious problem that calls fora proactive approach to comprehending, keeping an eye on, and reducing any possible hazards connected to the malicious application of generative AI in the context of malware creation. In the rapidly changing field of cybersecurity, addressing this problem is essential to creating efficient defences and guaranteeing theresponsible application of generative AI technology.

G. AI Powered Social -Engineering Attacks:

Generative AI has the potential to create complex social engineering schemes because of its capacity to imitate human-like interactions and produce persuasive information. This adds a new level of risk since bad actors could use content created by AI to trick people and coerce them into disclosing personal information or acting against their better judgement. In the context of social engineering concerns, it is critical tosimultaneously recognise and solve the growing issues that generative AI brings as we investigate its prospective uses. Creating strong awareness, education, and defences against the possible exploitation of generative AI in social engineering situations is essential for protecting people and organisations.

H. Deepfake Generation:

Generative AI has the potential to create synthetic media through the manipulation of existing content. This can include deepfakes, which are videos edited using artificial intelligence to replace or synthesize faces, speech, or emotions. Deepfakes pose risks if used withoutproper context or verification. They could spread misinformation about public issues andindividuals. Some concerns include using synthetic media to damage reputations, fabricate evidence, or mislead the public. Nation states may attempt to create confusion or erode trust in democratic institutions through manipulated videos. As media editing tools become more advanced, it is important for viewers to approach online content critically. Before sharing information widely, people should seek to confirm the authenticity and truthfulness of the material from reliable sources. Awareness of deepfake capabilities and vigilance in verificationcan help address the risks of manipulated media being used to spread falsehoods or undermine trust in traditional reporting.

The following is a case study of a popular youtuber and a social media influencer who became a victim of the threat of deepfake generation.

1.Case study:

TITLE: THE RISE OF DECEPTIVE DEEPFAKE ADVERTISING: A CASE STUDY ON MRBEAST'S TIKTOK INCIDENT.

INTRODUCTION:

The case study examines a recent incident on TikTok involving a fraudulentMrBeast advertisement that showcased the
increasing threat of deepfake technology in online marketing. Deepfake technology, driven by artificial intelligence, has become remarkably advanced, making it challenging for platforms like TikTok to detect and prevent deceptive content.

BACKGROUND:

MrBeast, a popular YouTube creator known for his extravagant acts of generosity, was targeted inthis incident. The deepfake advertisement offeredan iPhone 15 Pro at a price of just $2 to 10,000 viewers, a highly suspicious offer that would typically be recognized as a scam. However, given MrBeast's reputation for extraordinary giveaways, the deceptive ad appeared more convincing to unsuspecting users.[5]

KEY POINTS:

1) The Trust Factor:

MrBeast had built a strong reputation for his philanthropic stunts, such as giving away homes and cars without any conditions. Trust in MrBeast's authenticity made it easier for the deceptive deepfake ad to deceive users.

2) The Role of Deepfake Technology:

Deepfake technology has advanced to a point where it can create incredibly realistic content, making it difficult to distinguish between real and fake. The fraudulent MrBeast ad exploited this technology to impersonate the creator

3) TikTok's Moderation Challenges:

TikTok employs a combination of human moderation and AI-driven technology to review andapprove ads. In this case, TikTok's AI struggled to detect thefraudulent deepfake ad, allowing it to slip through the moderation process.

AREAS IMPACTED:

Deceptive deepfakes are not limited to MrBeast but also target other celebrities and public figures, fooling both younger and older audiences. Tom Hanks and Gayle King, two notable figures, have also fallen victim to deepfake advertisements.

REGULATORY CHALLENGES:

The Federal Trade Commission (FTC) has issued warnings about deepfake marketing, but regulating and preventing deceptive advertising at scale remains a significant challenge. The potential consequences of deceptive advertising, especially with globalelections approaching, raise concerns about theimpact on public perception and trust.

CONCLUSION:

The case of the fraudulent MrBeast ad on TikTokillustrates the growing threat of deepfake technology in online advertising. Creators with strong reputations can inadvertently aid scammers in deceiving users. Platforms like TikTok face challenges in detecting and preventing deepfake

content, emphasizing the need for continued advancements in AI-driven moderation. Additionally, regulatory bodies like the FTC must work to develop strategies to combat deceptive advertising, especially as the consequences of such practices become more severe in the context of elections and public trust.

IV. CONCLUSIONS

Generation AI (Gen AI) has enormous potential to revolutionize several sectors, but its rapid development also brings important cybersecurity and privacy issues. Cybersecurity threats Deepfakescan spread misinformation and undermine trust. Targeted phishing attacks can avoid detection. Automated malware development poses a significant threat. Privacy threats Data breaches and unauthorized access to personal data. The amplification of biases in data leads to discrimination. The grey GAI models hinder the understanding and identification of prejudices. Mitigation Strategies Robust cybersecurityframeworks with GAI-specific threat mitigation strategies. The Privacy-by-Design Principle incorporated data minimization, encryption and access control. Improve the transparency and explainability of the GAI model. Other measures include public awareness and education to empowerindividuals to protect data, ethical guidelines and regulatory frameworks to ensure appropriate use of Gen AI. By responding to these concerns, we can take advantage of GAI benefits while ensuring the security and privacy of our digital data.

ACKNOWLEDGMENT

We would like to convey our gratitude and extend our heartfelt appreciation to the HoD of the Cyber Security of the Shah and Anchor Kutchhi Engineering College Dr. Nilakshi Jain for providingus the support, essential resources and guidance for the completion of this survey paper. Her advice, guidance and encouragement have been crucial in forming this work's substance.

REFERENCES

[1] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng and F. -y. Wang, "Generative adversarial networks: Introduction and outlook", in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 588-598, 2017, doi: 10.1109/JAS.2017.7510583.

[2] R. Wei, C. Garcia, A. El-Sayed, V. Peterson and A. Mahmood, "Variations in Variational Autoencoders - A Comparative Evaluation," in IEEE Access, vol. 8, pp. 153651-153670, 2020, doi: 10.1109/ACCESS.2020.3018151.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Oord, Y. Bengio, and A. Courville, "Generative ai: A review of methods and applications," arXiv preprint arXiv:2309.07930, 2020.

[4] M. Gupta, C. Akiri, K. Aryal, E. Parker and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy," in IEEE Access, vol. 11, pp. 80218-80245, 2023, doi: 10.1109/ACCESS.2023.3300381.

[5] R. Pasupuleti, R. Vadapalli and C. Mader, "Cyber Security Issues and Challenges Related to Generative AI and ChatGPT," 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), Abu Dhabi, United Arab Emirates, 2023, pp. 1-5, doi: 10.1109/SNAMS60348.2023.10375472.

[6] A. Golda *et al*., "Privacy and Security Concerns in Generative AI: A Comprehensive Survey," in *IEEE Access*, vol. 12, pp. 48126-48144, 2024, doi: 10.1109/ACCESS.2024.3381611.

[7] S. Oh and T. Shon, "Cybersecurity Issues in Generative AI," 2023 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, Republic of, 2023, pp. 97-100, doi: 10.1109/PlatCon60102.2023.10255179.

[8] M. M. Rahman, A. Siddika Arshi, M. M. Hasan, S. Farzana Mishu, H. Shahriar and F. Wu, "Security Risk and Attacks in AI: A Survey of Security and Privacy," *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, Torino, Italy, 2023, pp. 1834-1839, doi: 10.1109/COMPSAC57700.2023.00284.

[9] Liu, Y., Huang, J., Li, Y. *et al.* Generative AI model privacy: a survey. *Artif Intell Rev* 58, 33 (2025). https://doi.org/10.1007/s10462-024-11024-6

[10] D. Zhang *et al*., "Privacy and Copyright Protection in Generative AI: A Lifecycle Perspective," *2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, Lisbon, Portugal, 2024, pp. 92-97.

[11] Nusrat Samia, Sajal Saha, Anwar Haque, Predicting and mitigating cyber threats through data mining and machine learning, Computer Communications, Volume 228, 2024, 107949, ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2024.107949.

[12] Y. Wang *et al*., "Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2245-2298, Fourthquarter 2023, doi: 10.1109/COMST.2023.3319492.

[13] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," in *IEEE Access*, vol. 10, pp. 36429-36463, 2022, doi: 10.1109/ACCESS.2022.3151903.

[14] Al-khazraji, Samer & Saleh, Hassan & Khalıd, Adil & Mishkhal, Israa. (2023). Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications.

[15] P. L. Kharvi, "Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media," in *IEEE Security & Privacy*, vol. 22, no. 4, pp. 115-122, July-Aug. 2024, doi: 10.1109/MSEC.2024.3405963