A comprehensive study on Generative AI as a double-edged sword in the digital security landscape

Dr. Purva Desai
M. B. Patel Science College, Anand, Gujarat, India.
Email:- dr.purva.desai@gmail.com

Abstract:

Generative Artificial Intelligence (GenAI) has emerged as a powerful technology enabling the autonomous production of highly realistic content in a variety of fields, including text, images, audio, and video. The ability of GenAI to learn and generate novel outputs alike human intellect has significantly influenced the global landscape, initiating a transformative era. However, this transformative potential has also been exploited by malicious actors to generate complex cyber threats. These include phishing emails, deepfake content, automated malware, and the dissemination of misinformation through seemingly credible social media posts, thereby introducing a novel array of challenges and risks within the field of cyber security. The necessity for strong countermeasures to safeguard digital infrastructures arises from these proficiencies. This review paper explores the dual-use characteristics of Generative AI in the realm of cyber security, offering a comprehensive analysis of the risks associated with its potential misuse. This study highlights gaps in existing countermeasures and identifies key areas for future research.

Introduction:

The rapid growth of Generative Artificial Intelligence (GenAI) is bringing major changes into innovation, revolutionizing creativity, problem-solving, and automation across various sectors. GenAI's ability to generate human-like text, realistic images, and deepfakes highlights its impressive potential to increase productivity and efficiency. However, this groundbreaking technology poses significant challenges in the area of cyber security. On one hand, GenAI has become a vital resource for fighting against cyber threats. Its expertise in analyzing large datasets, detecting anomalies, and predicting vulnerabilities in real time will be a notable advantage for organizations in securing their digital assets [1]. The automated threats identification and the smart defense frameworks aided by GenAI are expected to stay ahead of the ever-changing strategies employed by cybercriminals. On the other hand, the same attributes that position GenAI as a helpful instrument can also facilitate its use by cybercriminals [2]. The capacity to produce realistic fraudulent content, develop polymorphic malware, and exploiting zero-day vulnerabilities on a large scale poses substantial obstacles for cybersecurity professionals, challenges that were once deemed fictional [3]. Cybercriminals are increasingly utilizing Generative AI to craft phishing schemes that closely resemble legitimate communications, as well as to automate attacks that can adapt in real-time to defensive measures. This dual nature defines GenAI as a 'double-edged sword' in the context of cyber security. The ability to promote innovation, along with the possibility of misuse, raises important questions about trust, ethics, and the future of digital security. How can organizations effectively utilize GenAI to enhance their defensive measures while avoiding the unintended consequence of strengthening adversaries? Additionally, what importance do regulatory frameworks, ethical considerations, and international collaboration hold in addressing the risks associated with GenAI? This research paper addresses the dual characteristics of Generative AI in the field of cyber security, presenting a thorough analysis of its advantages, associated risks, and the actions needed to maintain a balance. By recognizing the opportunities it offers in conjunction with its associated risks, it intends to equip stakeholders with the necessary insights to effectively navigate this complex and rapidly evolving landscape [4].

COMPREHENSIVE ANALYSIS OF DUAL IMPACT

Generative Artificial Intelligence (GenAI) has significantly influenced cyber security, offering both innovative solutions and introducing new challenges. Recent literature provides a comprehensive overview of this dual impact:

1) Defensive Applications of GenAI

Generative Artificial Intelligence (GenAI) is progressively utilized to enhance cyber security measures. Its ability to process vast amounts of data, recognize patterns [5], and adapt dynamically to evolving threats makes it a critical asset in building resilient cyber security frameworks. The following section provides a comprehensive examination of the defensive roles of GenAI:

a) Threat Detection and Anomaly Identification: Automated threat detection is significantly improved through the application of Generative AI (GenAI), which analyzes extensive datasets to identify patterns indicative of potential threats. This includes monitoring network traffic for anomalies [5], such as unexpected spikes in data transfer or unauthorized access attempts. Additionally, GenAI employs behavioral analytics to assess user and system behaviors, establishing baseline norms. Any deviations from these established behaviors are flagged as potential indicators of insider threats, compromised accounts, or malware infections [6]. This dual approach enhances the overall security posture by proactively identifying and addressing threats before they can escalate.

b) Incident Response and Mitigation: The integration of Generative AI (GenAI) into Security Orchestration, Automation, and Response (SOAR) platforms significantly enhances automated incident response capabilities [7]. This technology streamlines repetitive and time-sensitive tasks, allowing for actions such as quarantining infected devices, blocking malicious IP addresses, and applying patches to vulnerable systems to be executed automatically. GenAI facilitates dynamic defense generation by enabling the development of real-time mitigation strategies that are specifically tailored to various attack vectors [8]. It can simulate potential attack scenarios, which allows for proactive recommendations regarding necessary configuration changes or the deployment of additional security controls. This proactive approach enhances overall security posture and responsiveness to emerging threats [8].

c) Threat Intelligence Enhancement: Intelligent Data Analysis through GenAI models enables the examination of both structured and unstructured data from various sources, including threat intelligence feeds, social media, and dark web forums [9]. This analysis results in actionable insights that can enhance security measures. GenAI's expertise in Language Translation and Contextual Understanding allows for the analysis of multilingual threat reports and social media discussions [10]. This capability is particularly important for global organizations, as it helps identify threats from different regions, thereby improving their overall threat detection and response strategies.

d) Dynamic Honeypots and Deception Technology: GenAI enhances traditional honeypots by making them dynamic and realistic, allowing them to mimic an organization's environment and adjust in real-time to deceive attackers. AI-generated fake documents, credentials, or databases can be strategically placed in networks to gather intelligence [11].

e) Post-Incident Analysis: Forensic analysis automation is enhanced by GenAI, which aids in post-incident investigations. It analyzes logs, reconstructs attack timelines, and identifies root causes [12]. Tools like Splunk Phantom [13] are used to correlate logs and security alerts, offering actionable forensic insights. It provides preemptive recommendations after incidents [14]. These suggestions aim to prevent future occurrences, including updates to firewall rules and adjustments to access control policies.

While GenAI holds immense promise for cybersecurity defense, organizations must address: (i) False Positives/Negatives: Balancing accuracy in detection to minimize disruptions [15]. (ii) Resource Dependency: Ensuring that smaller organizations can access affordable and scalable AI-powered tools. (iii) Adversarial Exploitation: Protecting GenAI systems themselves from being manipulated by attackers.

2) Offensive Exploitations of GenAI

Generative Artificial Intelligence (GenAI) is not only a powerful tool for defense but also a potent weapon in the hands of cybercriminals. Its ability to generate high-quality, context-aware content, automate complex processes, and adapt dynamically has amplified the sophistication and scale of cyberattacks [16]. Below is an in-depth exploration of the offensive applications of GenAI in cyber security:

a) Automated Phishing and Social Engineering: GenAI allows attackers to create hyper-personalized phishing emails by analyzing public and leaked datasets, mimicking legitimate communication styles. They can also use GenAI to translate phishing campaigns into multiple languages with high accuracy, broadening their scope. Malicious AI chatbots are deployed to impersonate customer service agents, soliciting sensitive information like login credentials or personal data.

b) Deepfake Technology for Cybercrime: GenAI-generated deepfakes impersonate high-level positions, authorizing fraudulent transactions or misleading stakeholders. Voice cloning for social engineering allows attackers to make convincing phone calls to authorize transactions or access sensitive information [17]. Disinformation campaigns use GenAI to create realistic fake news articles, social media posts, and propaganda, manipulating public opinion and creating chaos in organizations. These methods can lead to significant financial losses and disruptions in organizations.

c) Ransomware Evolution: GenAI is used in ransomware to implement adaptive encryption techniques, targeting specific file types or systems for maximum damage. Attackers use GenAI-powered bots to negotiate ransom payments, making conversations convincing and difficult to detect [18]. GenAI is also used for target selection and profiling, identifying high-value targets due to financial capacity or sensitive operations.

d) Automated Fraud and Scam Campaigns: GenAI is used for fake reviews, online manipulation, synthetic identities, and cryptocurrency scams. It generates fake product reviews, testimonials, and social media posts to manipulate public perception and conduct financial fraud. Attackers create synthetic identities with realistic names, photos, and credentials, which are used for fraudulent activities like opening bank accounts or credit cards.

e) Adversarial Attacks on AI Systems: Attackers generate malicious examples to confuse AI-driven cybersecurity systems, such as malware detection algorithms or biometric systems. GenAI is used to identify and exploit vulnerabilities in other AI systems [19], such as evading AI-based fraud detection in financial transactions.

Challenges in Addressing GenAI exploitation are (i) Accessibility: Open-access GenAI models lower the entry barrier for cybercriminals, enabling even low-skill attackers to execute sophisticated campaigns. (ii) Detection Difficulties: GenAI-generated content, such as phishing emails or deepfakes, is highly realistic, making it harder to detect and mitigate. (iii) Evolving Threats: The adaptability of GenAI allows attackers to continuously refine their tactics, keeping them one step ahead of traditional defenses.

3) Ethical and Security Challenges

GenAI introduces significant ethical and security challenges. These challenges arise from the dual-use nature of the technology, the rapid pace of its evolution, and the broader implications for privacy, trust, and regulatory governance. Below is a comprehensive exploration of these challenges:

a) Dual-Use Technology Dilemma: GenAI tools are dual-use, capable of both positive and negative purposes. Positive uses include threat detection and cyber resilience, while negative uses include malware generation and disinformation campaigns [20]. Open-access models like ChatGPT democratize AI capabilities but also empower malicious actors for cyberattacks. Restricting access to advanced models conflicts with inclusivity and innovation principles, raising ethical concerns about limiting AI access.

b) Ethical Concerns in Development and Deployment: GenAI models can perpetuate biases in their training data, leading to discriminatory outputs or uneven threat assessments. Accountability and liability are crucial, with developers, organizations, and end-users all responsible. Establishing clear accountability frameworks for damages caused by GenAI-generated outputs is an ethical dilemma. GenAI models often function as "black boxes," making it difficult to explain decisions or predictions [21], hindering trust and complicating ethical evaluations.

c) Challenges in Regulation and Governance: The global cybersecurity threat to GenAI is exacerbated by fragmented regulations, leading to inconsistencies in enforcement and application. Overregulation can stifle innovation and delay the development of beneficial applications, increasing the risk of malicious use and unchecked exploitation. GenAI systems' adaptability makes it difficult to effectively monitor and regulate misuse, and attackers often operate in jurisdictions with weak or no regulations, evading accountability.

d) Challenges in Mitigating Offensive Uses of GenAI: GenAI's automation capabilities enable realism and undetectability in phishing emails, deepfakes, and malware, making detection difficult. It also enables attackers to scale their operations, targeting thousands or millions simultaneously, such as automating spear-phishing emails for entire organizational hierarchies [21].

e) Resource Imbalance in Defense vs. Offense: GenAI reduces cybercriminal skill barriers, allowing low-skill actors to launch sophisticated attacks. However, defending against GenAI-driven attacks requires significant resources, disadvantaging small organizations.

f) Ethical Dilemmas in Countermeasures: GenAI's potential for deception raises ethical questions about its acceptableness for defensive purposes. Additionally, GenAI-driven monitoring systems can infringe on user privacy, posing concerns about surveillance and misuse of collected data.

g) Social and Psychological Impacts: GenAI-generated deepfakes and disinformation campaigns undermine trust in digital communications, public figures, and institutions [22], while attackers use GenAI to create manipulative content to intimidate, mislead, or psychologically harm individuals or groups.

h) Rapid Evolution of Threats: GenAI models rapidly evolve, outpacing regulatory and defensive frameworks' ability to adapt. Ethical concerns arise in keeping up with innovation while ensuring misuse protections remain effective [23]. Unintended consequences include attackers innovating countermeasures, creating an ongoing arms race with unpredictable outcomes.

4) Policy, Regulation, and Future Directions

Here’s a more detailed exploration, addressing the interplay between policy frameworks, technical advancements, ethical considerations, and long-term strategies in the context of Generative AI (GenAI) in cyber security:

a) Policy Considerations for GenAI in Cyber security: GenAI models face challenges in data privacy, intellectual property (IP), and liability. Large datasets for training often contain sensitive or personal data, necessitating policies for data anonymization, pseudonymization, and compliance with global data protection laws [24]. Unauthorized use of copyrighted materials could lead to IP disputes, necessitating guidelines for fair use and attribution. Liability and accountability are also crucial, requiring clear liability frameworks for AI-generated cybersecurity responses and human oversight in critical decision-making scenarios. These issues highlight the need for robust policies and regulations to ensure the safety and integrity of GenAI models.

b) Regulatory Measures for GenAI in Cyber security: various studies outline the development of cybersecurity standards for GenAI tools, promoting self-regulation and ethical AI guidelines. It also emphasizes continuous monitoring and reporting of GenAI tools to detect misuse and report incidents to authorities [25]. It also criminalizes GenAI use for malware development, phishing campaigns, and large-scale automated cyberattacks. Export controls on advanced GenAI technologies are enforced to prevent misuse by adversarial states or entities.

c) Future Directions for GenAI in Cyber security: The research focuses on developing GenAI models to detect and mitigate zero-day vulnerabilities and Advanced Persistent Threats. Public-private partnerships are encouraged to accelerate innovation and address risks.

Ethical AI practices include promote explainable AI (XAI) for cybersecurity and bias mitigation mechanisms. Education and workforce development involve training cybersecurity professionals with GenAI tools and methodologies [26]. Global coordination involves establishing international treaties and encouraging multi-national bodies like the United Nations and OECD to adopt global frameworks for regulating AI in security contexts.

RECOMMENDATIONS TO ADDRESS CHALLENGES AND ETHICAL DILEMMAS IN GENAI CYBER SECURITY

1) Responsible Development Practices: three strategies for ensuring the security of AI models: red-teaming AI models, bias auditing, and built-in safeguards. Red-teaming involves regular adversarial testing to identify vulnerabilities in AI systems, while bias auditing evaluates training datasets and model outputs for bias. Built-in safeguards involve robust filtering mechanisms to block harmful content generation, preventing GenAI tools from being used for phishing messages or malware. These strategies are implemented through dedicated teams, diverse datasets, and monitoring post-deployment use to detect misuse patterns. These measures aim to minimize discriminatory or unethical outcomes in AI decision-making.
2) Transparent Governance: The goal is to create Explainable AI (XAI) systems that make GenAI decisions interpretable and accountable, fostering trust and accountability in cybersecurity decision-making. Techniques like feature attribution can be used to explain decisions, while ethical standards are established to ensure responsible design and usage of GenAI tools. Implementation includes a Code of Conduct for AI developers and organizations, requiring documentation of ethical considerations throughout the AI development lifecycle.
3) Regulatory Collaboration: Global agreements on GenAI governance aim to ensure harmonized standards and enforcement across jurisdictions to address global cybersecurity threats. Implementation involves collaboration with multinational organizations to develop cross-border cooperation frameworks.
4) Public Awareness and Education: The initiative aims to raise awareness about GenAI risks and challenges, foster informed decision-making, and implement accessible resources on GenAI risks and best practices, while also launching public campaigns to highlight AI misuse prevention strategies.
5) AI-Driven Countermeasures: GenAI is being used to predict and mitigate cyberattacks in real-time, aiming to level the playing field against AI-driven threats. This involves

developing predictive models and autonomous AI agents to monitor and respond to network anomalies in real time. Additionally, Gen AI is being integrated into Security Operations Centers (SOCs) to enhance the ability to handle complex cybersecurity incidents, automating routine tasks and allowing analysts to focus on strategic decision-making with AI-supported insights.

CONCLUSION

Generative AI (GenAI) is a transformative technology that can significantly impact the cyber security landscape by enhancing defensive strategies, predicting attacks, and automating responses to threats. However, it also poses risks as cyber-criminals can exploit GenAI for malicious purposes, such as automating phishing campaigns and creating sophisticated malware. This dual-use nature raises ethical concerns and highlights the need for a multi-faceted approach to manage the associated risks. To tackle these challenges, a collaborative effort is needed among governments, organizations, academia, and AI developers to establish ethical guidelines, security standards, and regulations. Empowering cyber security professionals through continuous learning and fostering awareness at all levels is crucial to address GenAI risks effectively. By implementing safeguards and innovative approaches, GenAI can be a cornerstone for a secure digital future when used wisely.

REFERENCES

1) Prabhakar, S., Nalinaksha, I., & Anjaneyulu, V. (2023). Role of AI in enhancing cybersecurity measures to protect sensitive financial data. *International Journal of Science and Research Archive*. https://doi.org/10.30574/ijsra.2023.10.1.0700
2) Kumari, N., & Kumar, A. (2025). Advanced Computational Techniques for Analyzing Cybersecurity Event Datasets Using Artificial Intelligence and Machine Learning. *SCT Proceedings in Interdisciplinary Insights and Innovations.*, *3*, 524. https://doi.org/10.56294/piii2025524
3) Raji, A. N., Olawore, A. O., Ayodeji, A., & Joseph, J. (2024). Integrating Artificial Intelligence, machine learning, and data analytics in cybersecurity: A holistic approach to advanced threat detection and response. *World Journal Of Advanced Research and Reviews*. https://doi.org/10.30574/wjarr.2024.24.2.3197
4) Islam, S., Bari, Md. S., Sarkar, A., Khan, A., & Paul, R. (2024). AI-Powered Threat Intelligence: Revolutionizing Cybersecurity with Proactive Risk Management for Critical Sectors. *Deleted Journal*, *7*(01), 1–8. https://doi.org/10.60087/jaigs.v7i01.291

5) Kashyap, G. (2024). AI for Threat Detection and Mitigation: Using AI to identify and respond to cybersecurity threats in real-time. *Indian Scientific Journal Of Research In Engineering And Management*, *08*(12), 1–7. https://doi.org/10.55041/ijsrem10936

6) Gajiwala, C. (2024). Artificial Intelligence in Cybersecurity : Advancing Threat Modeling and Vulnerability Assessment. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. https://doi.org/10.32628/cseit241051066

7) Basu, S., Singh, U., Sharma, S., Kumar, G. A. S., & Upadhyay, R. (2024). *Generative AI Enabled Actionable Decision Support in Cyber Security Operations for Enterprise Security*. 1–8. https://doi.org/10.23919/ituk62727.2024.10772892

8) DM, V. G., & Ananda, A. (2024). Kecerdasan Buatan untuk Security Orchestration, Automation and Response: Tinjauan Cakupan. *Jurnal Komputer Terapan*, *10*(1), 36–47. https://doi.org/10.35143/jkt.v10i1.6247

9) Prieto, I., & Blakely, B. (2024). *Proposed Uses of Generative AI in a Cybersecurity-Focused Soar Agent*. https://doi.org/10.1609/aaaiss.v2i1.27704

10) Hayagreevan, H., & Khamaru, S. (2024). *Security of and by Generative AI platforms*. https://doi.org/10.48550/arxiv.2410.13899

11) Vast, R., Sawant, S., Thorbole, A., & Badgujar, V. (2021). Artificial Intelligence based Security Orchestration, Automation and Response System. *International Conference for Convergence for Technology*, 1–5. https://doi.org/10.1109/I2CT51068.2021.9418109

12) Al Adily, A. (2024). *Automating Incident Response with AI: Investigating how generative AI can streamline and automate incident response processes*. *06*(12), 569–575. https://doi.org/10.35629/5252-0612569575

13) Padamati, J., Nunnaguppala, L., & Sayyaparaju, K. (2021). Evolving Beyond Patching: A Framework for Continuous Vulnerability Management. Journal for Educators, Teachers and Trainers, 12(2), 185-193.

14) Silva, F. (2025). Navigating the dual-edged sword of generative AI in cybersecurity. *Brazilian Journal of Development*, *11*(1), e76869. https://doi.org/10.34117/bjdv11n1-062

15) Curran, K., Curran, E., Killen, J., & Duffy, C. (2024). The role of generative AI in cyber security. *Metaverse*, *5*(2), 2796. https://doi.org/10.54517/m.v5i2.2796

16) Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., & Isaac, W. (2024). Generative AI misuse: A taxonomy of tactics and insights from real-world data. arXiv preprint arXiv:2406.13843.

17) Usman, Y., Upadhyay, A., Gyawali, P., & Chataut, R. (2024). Is generative ai the next tactical cyber weapon for threat actors? unforeseen implications of ai generated cyber attacks. arXiv preprint arXiv:2408.12806.

18) da Silva, F. A. (2025). Navigating the dual-edged sword of generative AI in cybersecurity. Brazilian Journal of Development, 11(1), e76869-e76869.

19) Krishnamurthy, O. (2023). Enhancing Cyber Security Enhancement Through Generative AI. International Journal of Universal Science and Engineering, 9(1), 35-50.

20) Huang, K., Goertzel, B., Wu, D., & Xie, A. (2024). GenAI model security. In Generative AI Security: Theories and Practices (pp. 163-198). Cham: Springer Nature Switzerland.

21) Tomassi, A. (2024). Data Security and Privacy Concerns for Generative AI Platforms (Doctoral dissertation, Politecnico di Torino).

22) Ruparel, H., Daftary, H., Singhai, V., & Kumar, P. (2024, December). The Impact of Generative AI on Cloud Data Security: A Systematic Study of Opportunities and Challenges. In 2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing (UCC) (pp. 185-188). IEEE.

23) Verma, A., Sankhyayan, S., Jawanda, K., & Tandon, S. (2024, August). Generative Artificial Intelligence and Cybersecurity Risks: Issues and Challenges. In International Conference on ICT for Sustainable Development (pp. 321-327). Singapore: Springer Nature Singapore.

24) Christodorescu, M., Craven, R., Feizi, S., Gong, N., Hoffmann, M., Jha, S., & Turek, M. (2024). Securing the future of GenAI: Policy and technology. arXiv preprint arXiv:2407.12999.

25) Huang, K., Joshi, A., Dun, S., & Hamilton, N. (2024). AI regulations. In Generative AI security: theories and practices (pp. 61-98). Cham: Springer Nature Switzerland.

26) Shaznay, N. (2025). Bridging the Industry-Higher Education Gap with Critical Design Futures Thinking and GenAI for Innovation. In The Rise of Intelligent Machines (pp. 268-288). Chapman and Hall/CRC.