Chapter 9

Dark Side of Generative AI: Emerging Threats and Attack Vectors

[1]Umar Farooq, [1]Mina Bakhtiyar Ahmad Khan, [2]Kounser Ali Mir, [1]Parvinder Singh, [1]Surinder Singh Khurana, [1]Anam Bansal
[1]Dept. of Computer Science & Technology, Central University of Punjab, Bathinda, India
[2]Dept. of Computer Science, Akal University, Talwandi Sabo, India

Abstract

In today's digital age, generative AI is transforming various sectors. But its dark side is becoming increasingly evident, revealing a growing range of threats and attack vectors. Cybercriminals use it to spread harmful content, especially targeting women and children, which harms trust in society and worsens social problems like online abuse. It motivates us to study its emerging threats, attack vectors, and the corresponding case studies. It is used to manipulate information, breach security, and conduct advanced cyber scams. It is used to craft convincing phishing emails and messages to trick people more easily and extort money. It also enables the automated creation of malicious code, disinformation campaigns, and hallucinated content, amplifying both the speed and scale of cybercrime. It enables unethical practices such as political deepfakes, identity theft, vishing, and the creation of deepfake pornography used for revenge or reputational damage. Alarmingly, the rise of dark LLMs designed explicitly for malicious purposes represents a new frontier in cybercrime. Given the growing threats, there is an urgent need for increased public awareness, strong policy frameworks, and advanced mitigation strategies. Addressing these challenges is essential to prevent misuse while preserving its benefits in society across healthcare, education, business, and more.

Keywords. Generative AI, deepfake, cybersecurity, vishing, malware, social engineering, hallucination, disinformation.

1 Introduction

1.1 Rise of Generative AI

Artificial intelligence (AI) has rapidly evolved from rule-based systems to machine learning. Traditional AI systems are limited to data analysis, pattern recognition, prediction, and automation, while generative AI advances to create new content [1]. It can generate human-like text, realistic images and videos, original music, 3D environments, scientific formulas, and complex computer code. The rise of generative AI can be attributed to advancements in machine and deep learning, natural language processing, access to large datasets, increased computational power, and open-source contributions [2]. Today, generative AI is one of the most groundbreaking advancements in AI. As shown in Figure 1a, the state-of-the-art architectures behind this revolution include, but are not limited to, the following [1]:

- Generative Adversarial Networks (GANs) contain two neural networks, the generator and the discriminator, that work within a loop. The generator tries to generate new content (like images or videos) while the discriminator evaluates whether it looks real or fake. Over time, both networks improve, resulting in highly realistic outputs [3]. It is used in StyleGAN and Artbreeder to create highly realistic images, often of human faces.