IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 1, January 2025

DOI: 10.17148/IARJSET.2025.12111

The Dark Side of AI: How Cybercriminals Are Weaponizing Machine Learning

Enoch Anbu Arasu Ponnuswamy

Barclays Bank, New Jersey, USA

Abstract: Generative AI is revolutionizing the way industries operate with its positive impact on patient care, claims processing, and customer service being the most recognizable. However, alongside this advancement, these is a darker side to AI that many overlook.

And as technology becomes more integral to our daily lives, cybercriminals are increasingly leveraging artificial intelligence (AI) to carry out highly sophisticated attacks, posing serious risks to both individuals and organizations. Recent high-profile data breaches, such as the one involving Star Health which led to personal data of 31 million customers being compromised underscore the growing danger of this new threat landscape. In this article, we'll explore the various types of attacks being carried out by cybercriminals using AI, shedding light on the growing threats they pose.

Keywords: Artificial Intelligence (AI), cybercriminals, Generative AI, Multi-Factor Authentication (MFA).

I. INTRODUCTION

Automated Phishing Campaigns

Phishing attacks have existed since the early days of the internet. These attacks that prey on the kindness of human nature have evolved from the days of Nigerian princes seeking financial aid and rich relatives looking for beneficiaries to increasingly sophisticated attacks that impersonate customer support agents from reputed banks and even government owned entities.

And this has only been made worse with the rise of Gen AI. Cybercriminals are now using AI algorithms to analyze vast amounts of data, such as social media profiles and publicly available information, to craft highly personalized phishing emails. These messages appear more legitimate and are harder to detect as fraudulent, increasing their success rate.

For instance, AI can analyze a target's communication style, preferred topics, and typical email formatting to create phishing emails that mimic a trusted colleague or service provider. This personalization significantly increases the likelihood of a victim clicking on a malicious link or divulging sensitive information.

And what's worse since the fourth quarter of 2022 (which was around when ChatGPT burst onto the scene), there's been a 1,265% increase in malicious phishing emails,

Mitigation Strategies:

- **AI-Powered Email Filters:** Deploy advanced spam filters that use machine learning to detect subtle signs of phishing, such as unusual sender behavior or mismatched URLs.
- **User Training and Awareness:** Conduct regular training sessions to teach users how to identify phishing attempts, even those crafted with AI.
- **Multi-Factor Authentication (MFA):** Require MFA to add an extra layer of security, even if credentials are compromised through phishing.
- **Real-Time Phishing Detection:** Implement browser-based anti-phishing tools that alert users when visiting malicious websites.

Malware Powered by Machine Learning

Machine learning algorithms enable malicious software to adapt to its environment, evade detection, and bypass traditional security measures. For example, polymorphic malware can alter its code each time it's executed, making it nearly impossible for signature-based antivirus systems to identify.

Even worse, AI can optimize malware's behavior. It can learn how to avoid sandbox environments, detect patterns in user activity, and decide the best time to execute its payload for maximum damage. This adaptability gives AI-driven malware a significant edge over traditional threats.

Mitigation Strategies:

- **Behavior-Based Detection:** Use AI-driven endpoint detection and response (EDR) systems that analyze behavior patterns to detect polymorphic malware.
- **Threat Intelligence Sharing:** Participate in industry-wide threat intelligence networks to stay informed about evolving AI malware trends.
- **Regular Updates and Patching:** Ensure all software and systems are up-to-date to reduce vulnerabilities that AI-driven malware can exploit.
- **Network Segmentation:** Limit the spread of malware by segmenting networks and enforcing strict access controls.

II. LITERATURE REVIEW

Deepfakes and AI-Generated Fraud

According to Deloitte, a leading financial research group, AI-generated content contributed to more than $12 billion in fraud losses last year. Deepfakes use AI to manipulate video and audio, creating hyper-realistic media that can impersonate individuals with stunning accuracy. This technology has already been weaponized in cyberattacks to commit fraud and deceive targets.

One prominent example of AI-driven fraud is the deepfake video featuring Elon Musk, which was used to deceive and defraud unsuspecting victims. Steve Beauchamp, an 82-year-old retiree, was one of those victims.

Late last year, Beauchamp watched a video of Musk endorsing a high-risk investment opportunity that promised rapid returns. Believing it was a legitimate chance to secure his family's financial future, he contacted the company behind the pitch and opened an account with $248. Over several weeks, Beauchamp drained his retirement savings, ultimately investing more than $690,000.

But then, the money vanished—lost to digital scammers at the forefront of a new criminal enterprise powered by artificial intelligence. The scammers had used AI to create a convincing deepfake of Elon Musk. By manipulating a genuine interview, they replaced his voice with an AI-generated replica and adjusted his mouth movements to match the altered script. To a casual viewer, the manipulation was nearly imperceptible, showcasing the alarming sophistication of this new form of digital fraud.

III. PROPOSED MODEL

Mitigation Strategies:

- **Verification Protocols:** Require multi-channel verification for sensitive requests, such as confirming transactions via video call or written approval.
- **AI Deepfake Detection Tools:** Use AI tools specifically designed to identify manipulated audio or video content, such as Microsoft's Video Authenticator or Sensity.
- **Employee Education:** Train employees to spot deepfake inconsistencies, like unnatural facial movements or mismatched voice tones.
- **Media Watermarking:** Implement digital watermarks in official videos and audio to verify authenticity.

Social Engineering

By creating realistic fake profiles and crafting convincing messages, AI allows attackers to simulate genuine interactions that help them gain the trust of victims. Cybercriminals can use these tools to generate phishing emails, text messages, or even fake social media conversations that appear authentic and urgent.

These fake identities can be designed to mirror people or organizations the victim already knows, making it easier for criminals to deceive their targets and allows attackers to infiltrate networks, steal sensitive information, or manipulate individuals into taking harmful actions

Whether it's impersonating a colleague, a romantic interest, or a customer support agent, AI makes it possible for attackers to manipulate victims over time, building rapport and increasing the chances of success. As AI continues to evolve, its role in social engineering attacks will only grow, making these scams more effective and harder to detect.

Mitigation Strategies:

- **Zero Trust Architecture:** Implement a zero-trust security model that requires continuous verification of all users, devices, and systems accessing a network, regardless of their location.
- **Encryption:** Encrypt sensitive data both at rest and in transit using strong encryption algorithms to prevent unauthorized access.
- **Monitoring and Logging:** Enhance logging systems to capture detailed activity records, enabling faster detection of AI-powered threats.

DOI: 10.17148/IARJSET.2025.12111

- **Public Awareness Campaigns:** Launch campaigns to educate the public on identifying AI-based scams and maintaining digital hygiene.

IV. CONCLUSION

AI-driven cyber threats represent a significant and evolving challenge, but they are not insurmountable. By implementing robust cybersecurity practices, fostering education and awareness, we can counteract the misuse of gen AI and reduce the risks associated with these emerging threats.

REFERENCES

[1]. (https://www.cbsnews.com/texas/news/deepfakes-ai-fraud-elon-musk/)
By Brian New, Lexi Salazar, Mike Lozano, Scott Fralicks,Updated on: November 24, 2024 / 2:28 PM CST / CBS Texas
[2].(https://krebsonsecurity.com/2023/08/meet-the-brains-behind-the-malware-friendly-ai-chat-service-wormgpt/)
"Meet the Brains Behind the Malware-Friendly AI Chat Service 'WormGPT'" was authored by Brian Krebs and published on August 8, 2023.
[3].(https://www.indiatoday.in/technology/features/story/star-health-insurance-hack-led-to-personal-data-of-31-million-customers-being-compromised-story-in-5-points-2615354-2024-10-11)
[4]. Star Health insurance hack led to personal data of 31 million customers being compromised: Story in 5 points
Published on October 11, 2024, on India Today's website
[5].(https://www.deloitte.com/middle-east/en/our-thinking/mepov-magazine/securing-the-future/ai-in-cybersecurity.html) AI in cybersecurity: A double-edged sword published on Deloitte's Middle East website
[6].(https://cybermagazine.com/articles/cybercriminals-are-creating-a-darker-side-to-ai)
Cybercriminals are creating a darker side to AI published on October 24, 2023, in Cyber Magazine, was written by Katy Allan.