IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

GENERATIVE AI IMPLICATIONS FOR CYBER SECURITY

Vivaan Vavilala
Student
12th Grade
Chrysalis High Kadugodi, Bangalore, India

1 Abstract

This paper provides an overview of the current cybersecurity market, tools and the current state of cyberattacks. It shows the impact of cybercrimes not just on corporations but also on common people. It shows how machine learning is currently being used to defend against cyberthreats. It then delves into Generative AI and its use by both cyberattackers and cyberdefenders. It finishes with a look at ways to protect the GenAI infrastructure itself and provides caution on reliance on GenAI for cybersecurity.

2 Index Terms

Generative AI, Artificial Intelligence, Cyber Security, Cyber-Attacks, Cyber Safety

3 Current State of Cyber Security

Cyber security refers to the act of protecting computer systems, data, and networks from unauthorized access and attacks. It is estimated that roughly 54 people fall victim to a cyber-attack per second (Martin) and is projected to cost the world over $1.5 trillion by the end of 2025 (Milliefsky). The cybersecurity tools market to prevent cyberattacks is estimated to be around $190 billion in 2024 and is expected to grow to over $500 billion by 2032. (“Cybersecurity Market Size, Share, Analysis | Global Report 2032”). Thus, it is an arms race between the cyber attackers and the cyber defenders.

3.1 Current Cyber Attacks

Cyber-attack is an all-encompassing name given to any attack that targets the digital infrastructure through the network. There are several types of cyberattacks and techniques like

- Ransomware
- Denial of Service
- Data breach
- Phishing
- Social Engineering
- Malware
- Man-in-the-middle

- SQL injection
- Zero-day exploits

Each of these types of attacks have their own impact on the victim. While some ransomware result in payment, denial of service results in business outages, data breaches may result in leaking confidential or embarrassing information and so on. Research (Eviden) shows that ransomware attacks are on the rise

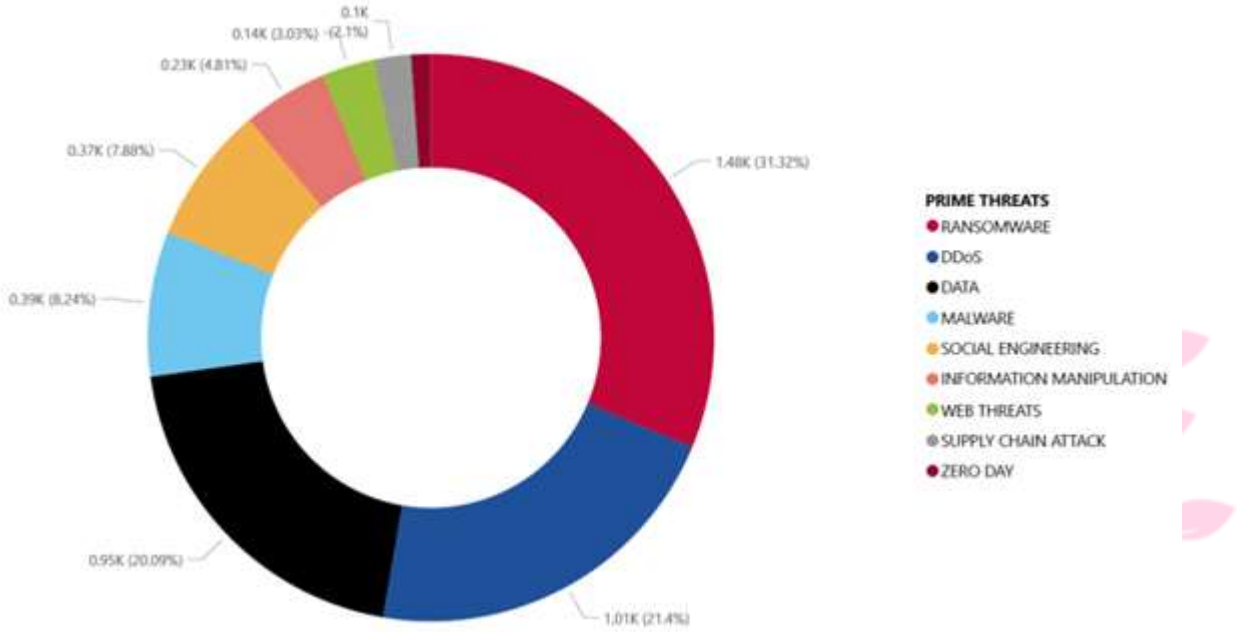

figure 1 distribution of various cyber-attack types in 2023

4 Is Cybercrime a victimless crime?

Cybercrime has been thought of as a victimless crime, since it usually involves digital assets and most of the losses that are publicized are to corporations which must pay off ransomware criminals or deal with data breaches. However, the scepter of cybercrime has changed. Cyber criminals have targeted individuals in the recent past. Investment fraud and romance scams top the list of crimes against individuals (Griffiths), resulting in losses over $2 billion. People over the age of 60 are disproportionately targeted as victims of cybercrime.

Cybercrime does stop at just financial implications. While direct stats are hard to come by, there are several cases of deaths or serious injury attributable to cybercrimes. In one case in Dusseldorf, Germany, a patient being taken to a hospital had to be redirected to another far away hospital due to a ransomware attack which caused his death on the way (Horne et al.). A 2023 paper (McGlave et al.) by researchers at the University of Minnesota's School of Public Health found that between 2016 and 2021, between 42 and 67 US Medicare patients died as a result of ransomware incidents. A retired couple in Belagavi died by suicide after allegedly being harassed and blackmailed by cybercriminals who impersonated law enforcement and extorted a large sum of money from them. (Vasudev). From these examples, it is obvious that victims or cybercrimes are not just corporations but real people as well.

5 Use of AI in Cyber security

Cybersecurity companies have been using AI tools to monitor, detect and prevent attacks for several years (Fitzgerald). Several use cases like the ones outlined below are already being used by corporations.

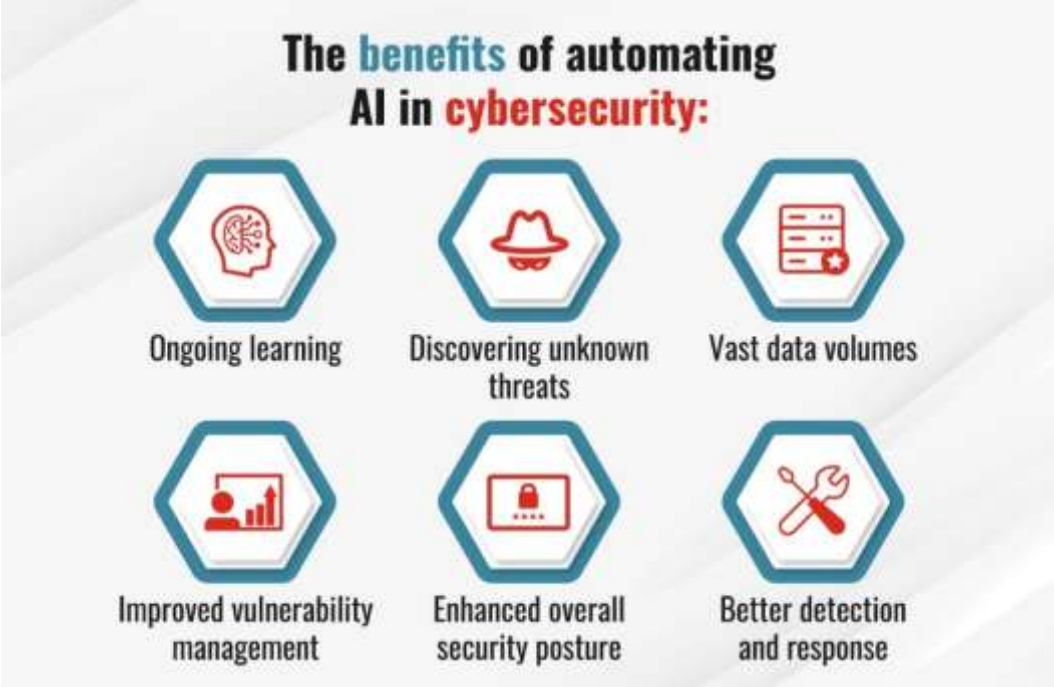

figure 2 use of ai in cybersecurity (Fortinet, n.d.)

5.1 Threat detection

This has been one of the primary uses of machine learning and anomaly detection. Prior to use of machine learning, legacy systems relied on signatures that were hand coded by security researchers. AI powered systems can detect patterns and anomalies without human intervention, which allows them to detect zero-day attacks. Anomalies from normal baselined behavior like an unknown user logging into a system or many connections to a particular service can become a trigger to detect a threat.

5.2 SPAM protection

SPAM email is one of the common complaints of all email providers as well as end users. Public email providers like Gmail receive approximately 15 billion spam emails every day. AI enabled SPAM detection tools today are effective at detecting and blocking over 99.9% of SPAM emails, thus ensuring that only a small percentage of SPAM ends up in the user's Inbox.

5.3 Fraud Detection

Machine learning and anomaly detection finds great use in fraud detection by financial institutions. ML models will model a pattern of user spending behavior and any anomaly for that will trigger an alert. For example, a transaction from a foreign country that the user has never travelled to or a series of rapid, small transactions will trigger an alert.

5.4 Threat Response

Threat response could include notifying administrators of an ongoing security incident or taking proactive remediation. In most deployed systems, notification is through an altering mechanism from a SIEM or a custom SOC dashboard. Then the administrators will act to mitigate and avert the threat. In some cases, the cybersecurity tools allow the tool to take proactive action automatically without human intervention using automated policies. This could include activating firewall rules or shutting down certain network services while the incidents are investigated.

6 Generative AI

Generative AI has taken the world by storm in the last couple of years. The applications of Generative AI are myriad, but from a cyber security perspective, it is a mixed blessing. It provides tools for both the attackers and defenders, triggering another arms race.

6.1 How GenAI is Used by Cyber Attackers

Generative AI provides new tools and attack vectors for the attackers to perpetrate cyber-crimes. These include the following:

6.1.1 Hyper personalized Social Engineering and Phishing

One of the ways that tools detected phishing was to look for grammatical errors, malformed or out of place sentences etc. primarily because the attackers generally are not in the same geo location as the victims. With Generative AI, attackers can generate perfect and personalized social engineering and spear phishing emails and messages which makes it harder for traditional tools to detect phishing.

6.1.2 Deepfakes

Realistic generation of deep-fake audio and video has resulted in cybercrimes where impersonation has become easy. ("What Are the Risks and Benefits of Artificial Intelligence (AI) in Cybersecurity?") In a recent incident (Chen and Magramo), a finance worker transferred $25 million after having a realistic video conversation with his Chief Financial Officer which turned out to be a deepfake. The quality of deepfakes generated by AI makes them very hard to discern for average human beings

figure 3 deepfake uses by cyberattackers (Fortinet, n.d.)

6.1.3 Automated Malicious tools

GenAI makes it easy to generate tools and botnets using the code authoring tools. It is very easy to make self-morphing worms, autonomous malware etc., which allows the attackers to overwhelm the defenses of normal cyber security tools.

6.2 How GenAI is Used by Cyber Defenders

The same GenAi that has become a force multiplier for cyber attackers has also enabled cyber defenders to be much more effective as well. (Sangfor) (SentinelOne)

6.2.1 Vulnerability Management and risk assessment

GenAI can read through vast amounts of code and assess the vulnerable code and suggest fixes. It can also look at various databases like CVE database and provide an assessment of the risk of those vulnerabilities

6.2.2 Advanced Threat Detection

GenAI can process huge amounts of logs and unstructured data sources to identify anomalies, which will help identify threats. While traditional machine learning models relied on structured data which required a lot of curated data, GenAI can take unstructured data and processing it to identify anomalies.

6.2.3 Endpoint protection

When GenAI is deployed in an endpoint protection solution, it allows an easy way to identify malware as the text is typically in an unencrypted form in the endpoints. This allows rapid detection and alerting of potential issues before they become more widespread.

6.2.4 Natural Language Querying

GenAI by its nature allows administrators to provide prompts in a natural language and query for information. This allows security researchers to be much more efficient and productive in identifying underlying threats without having to learn the syntax of underlying tools

6.2.5 Co-Pilots and Agentic AI

Given the spread of cyber security tools, it becomes very hard for security professionals to learn and monitor all the tools in a security operation center. It is generally understood that a medium sized enterprise deploys about 60-70 cybersecurity tools at any given time (Zorz). This is known as tool bloat and overwhelms security professionals. With the advent of GenAI based co-pilots and AI agents, the task of trawling through multiple tools and correlating data across tools can be simplified. While it doesn't reduce the tool bloat, it helps manage the impact of tool bloat much better

7 How to Protect GenAI Infrastructure from Cyber Attacks

While GenAI can be used as a tool by both cyber attackers and cyber defenders, the advent of GenAI has created a new attack vector: the GenAI infrastructure itself. There is now a need to protect the GenAI infrastructure from attacks.

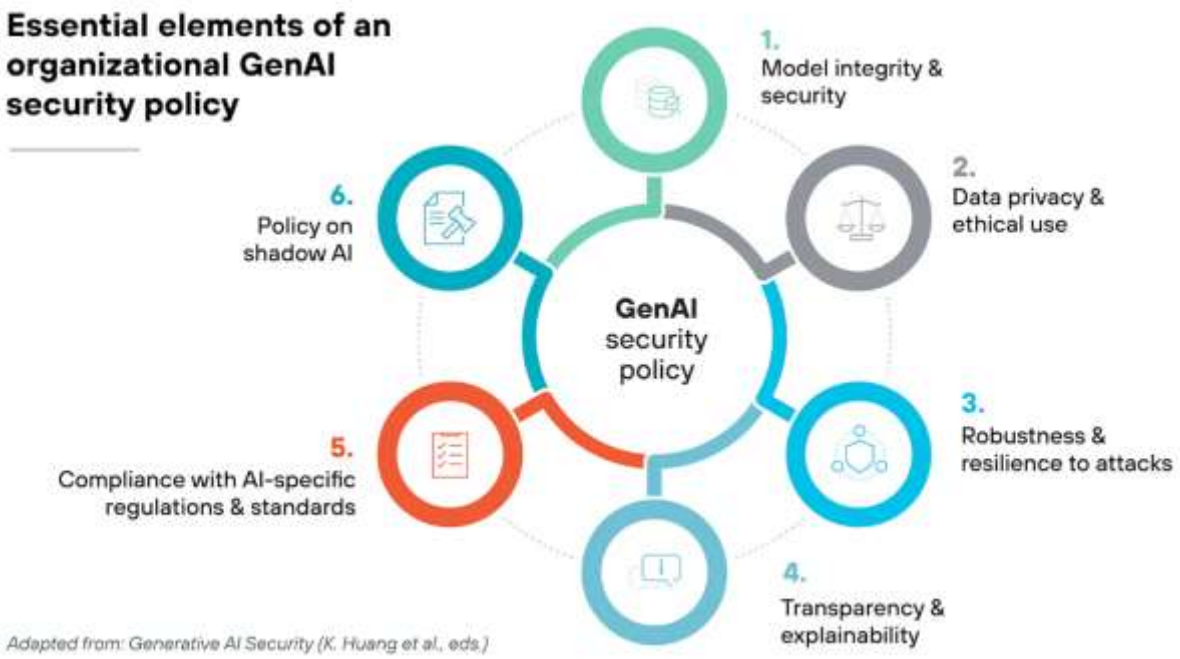

figure 4 genai based security policy (Palo Alto Networks Inc, n.d.)

7.1 GenAI Infrastructure Overview

The GenAI infrastructure consists of the following elements

7.1.1 Compute Infrastructure

GenAI compute data centres are distinct from traditional data centers. They tend to have much more concentration of GPUs and AI focused CPUs rather than general purpose servers which have been hardened against cyber-attacks over the years

7.1.2 Data pipeline

Data is a vital part of GenAI infrastructure. Protecting the data that models are trained on is vital. Data processing frameworks that are used to preprocess and sanitize data are also a key part of this pipeline.

7.1.3 AI/ML Models

Models are the brains of the GenAI infrastructure. They determine how data is processed, how decisions are made etc. This includes not just the training part but also MLOps, which involves deploying these models on compute resources, scaling them and updating them periodically.

7.1.4 APIs and Integrations

This is the final point where applications and users interact with the AI models for inference. This includes APIs, user interface, 3rd party applications (like SIEM, EDR) which leverage the models.

7.2 New Attack Vectors

Given the above components of AI infrastructure, new attack vectors are possible that will undermine how GenAI can be used to combat cybercrime

7.2.1 Prompt Injection

This is the most common attack vector in GenAI. When prompts can be surreptitiously entered into the GenAI without proper sanitization, it can wreak havoc on the model and the results. Several attacks like Echoleak (CVE-2025-32711) ("What Is a Prompt Injection Attack? [Examples & Prevention]") in the recent past have been successful leveraging this approach.

7.2.2 Model Poisoning

If the attacker can poison the model itself, then the inference results will end up with the wrong results even with the right data and prompts

7.2.3 Data Poisoning

Another way to attack the AI models is to poison the input data, especially data that is used to train the model. Most models are periodically trained on user data to continuously improve the results. By continuously feeding wrong data, the model can be trained to provide wrong results for future inference

7.2.4 Model theft

Models represent the intellectual property of the developer. It takes a huge amount of compute and human resources to successfully train a model. Several methods exist to steal the model or distill the contents of the model. This can be prevented by appropriate guard rails in protecting the model and its usage.

8 Other considerations when deploying GenAI for cyber security

In addition to the various potential adversarial attacks on GenAI infrastructure, there are other considerations where we need to be careful about using GenAI in cyber security.

8.1 Hallucination

As is well known, GenAI tends to hallucinate and provide outputs that don't necessarily match the data it is trained on. This can lead to potential wrong actions, if the inferences are not vetted by either a human or using RAGs (Retrieval Augmented Generation)

8.2 Model bias

While it may seem like models shouldn't have any biases unlike humans, model bias is a very common problem. Models are tuned by humans and hence will reflect the (Sekiri and Ainsworth) biases of the developers as well as the data that is fed into them. Examples include AI recruiting tools showing a bias towards men, racial bias in risk prediction algorithms etc.

8.3 Overreliance and knowledge gap

With the employee base not fully trained on GenAI, it is possible to develop an over-reliance on the inferences by GenAI tools ("What Are the Risks and Benefits of Artificial Intelligence (AI) in Cybersecurity?"). It reduces the vigilance and critical thinking provided by humans, allowing new attacks to bypass security issues that the models haven't been trained on. Given that there is an acute shortage of GenAI knowledgeable workforce, this will continue to be a problem without widespread retraining of employees

8.4 Regulatory Compliance

GenAI involves the use of large amounts of data. These data can exist across multiple regulatory boundaries (like EU vs USA). Different regulatory frameworks have different requirements for data retention, transport and storage. There is no easy way to force a GenAI model to forget data that it is trained on. Hence organizations need to be careful that the data usage follows the regulatory framework. Added to that, the regulatory framework on GenAI is nascent and evolving, which requires organizations to continuously monitor new regulations.

9 Summary

GenAI is a very promising technology for cyber defenders. Use of it requires caution and as well speed to stay ahead of the cyber attackers, who are using the same technologies. In addition, cyber defenders need to be aware of pitfalls of using GenAI without understanding its implications and take steps to protect their critical GenAI infrastructure.

References

Chen, H., & Magramo, K. (2024, February 4). *Finance worker pays out $25 million after video call with deepfake 'chief financial officer'.* Retrieved August 23, 2025 from CNN: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Cyber Attack - What Are Common Cyberthreats? (n.d.). Retrieved July 31, 2025 from Cisco: https://www.cisco.com/c/en_in/products/security/common-cyberattacks.html#~types-of-cyber-attacks

Cyber Management Alliance. (2025, Jan 20). *Top 10 Biggest Cyber Attacks of 2024 & 25 Other Attacks to Know About!* From Cyber Management Alliance: https://www.cm-alliance.com/cybersecurity-blog/top-10-biggest-cyber-attacks-of-2024-25-other-attacks-to-know-about

Cybersecurity Market Size, Share, Analysis | Global Report 2032. (2025, July 28). Retrieved August 18, 2025 from Fortune Business Insights: https://www.fortunebusinessinsights.com/industry-reports/cyber-security-market-101165

Eviden. (n.d.). *Trends in OT Security.* From Make your OT Security Come True in 2024: https://eviden.com/publications/digital-security-magazine/cybersecurity-predictions-2024/trends-in-operational-technology-ot-security/

Fitzgerald, A. (2024, May 6). *AI in Cybersecurity: How It's Used + 8 Latest Developments.* Retrieved August 23, 2025 from Secureframe: https://secureframe.com/blog/ai-in-cybersecurity

Fortinet. (n.d.). *Artificial Intelligence in CyberSecurity*. From https://www.fortinet.com/resources/cyberglossary/artificial-intelligence-in-cybersecurity

Fortinet. (n.d.). *What is a deepfake*. From Fortinet Cyberglossary: https://www.fortinet.com/resources/cyberglossary/deepfake

Griffiths, C. (2025). *The Latest Cyber Crime Statistics (updated July 2025) | AAG IT Support.* Retrieved August 23, 2025 from AAG IT Services: https://aag-it.com/the-latest-cyber-crime-statistics/

Horne, S., Mott, D., & MacColl, J. (2024, June 25). *Ransomware: A life and death form of cybercrime.* From rusi.org: https://www.rusi.org/explore-our-research/publications/commentary/ransomware-life-and-death-form-cybercrime

Martin, J. (2025, June 6). *How Many Cyber Attacks Occur Each Day? (2025).* Retrieved August 11, 2025 from Exploding Topics: https://explodingtopics.com/blog/cybersecurity-stats

McGlave, C. C., Neprash, H., & Nikpay, S. (2023, Oct 4). *Hacked to Pieces? The Effects of Ransomware Attacks on Hospitals and Patients.* From https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4579292

Milliefsky, G. (2025, March 13). The True Cost of Cybercrime: Why Global Damages Could Reach $1.2 – $1.5 Trillion by End of Year 2025. *Cyber Defense Magazine*. From https://www.cyberdefensemagazine.com/the-true-cost-of-cybercrime-why-global-damages-could-reach-1-2-1-5-trillion-by-end-of-year-2025/

Palo Alto Networks Inc. (n.d.). *How to Build a Generative AI Security Policy*. From PaloAlto Networks Cyberpedia: https://www.paloaltonetworks.com/cyberpedia/ai-security-policy

Sangfor. (2024, 03 06). *Generative AI in cyber security.* From Sangfor: https://www.sangfor.com/blog/cybersecurity/what-is-generative-ai-cybersecurity

Sekiri, N., & Ainsworth, O. (2025, April 29). *5 Real-life Examples of AI Bias.* Retrieved August 23, 2025 from Digital Adoption: https://www.digital-adoption.com/ai-bias-examples/

SentinelOne. (2025, 05 26). From What is Generative AI in CyberSecurity: https://www.sentinelone.com/cybersecurity-101/data-and-ai/generative-ai-cybersecurity/

Vasudev, A. (2025, March 29). *Karnataka: Elderly Couple In Belagavi Die by Suicide After Falling Victim To 'Digital Arrest' Scam Read more at: https://www.oneindia.com/bengaluru/karnataka-elderly-couple-in-belagavi-die-by-suicide-after-falling-victim-to-digital-arrest-scam-4108615.htm.* From oneindia.com: https://www.oneindia.com/bengaluru/karnataka-elderly-couple-in-belagavi-die-by-suicide-after-falling-victim-to-digital-arrest-scam-4108615.html

What Are the Risks and Benefits of Artificial Intelligence (AI) in Cybersecurity? (n.d.). Retrieved August 23, 2025 from Palo Alto Networks: https://www.paloaltonetworks.com/cyberpedia/ai-risks-and-benefits-in-cybersecurity

What Is a Prompt Injection Attack? [Examples & Prevention]. (n.d.). Retrieved August 23, 2025 from Palo Alto Networks: https://www.paloaltonetworks.com/cyberpedia/what-is-a-prompt-injection-attack

Zorz, M. (2025, March 27). *The hidden costs of security tool bloat and how to fix it.* From Helpnet Security: https://www.helpnetsecurity.com/2025/03/27/shane-buckley-gigamon-deep-observability-tool-stacks/