Central Queensland university (CQU) Intellectual property

COIT20265: NETWORK AND INFORMATION SECURITY PROJECT (HT2, 2024)

FINAL REPORT

PROJECT:

GENERATIVE AI: NAVIGATING SHORT-TERM SKEPTISM AND LONG-TERM PROMISE

UNDER THE ESTEEMED GUIDANCE OF

UNIT CO-ORDINATOR: FARIZA SABRINA

MENTOR: DR. AHMEDI AZRA

TEAM MEMBERS:

BASANTA ADHIKARI (12211752)

BHUWAN THAPA (12196590)

KIRAN BHUSAL (12211570)

PRATIK SHINGH DHAMI (12209929)

Table of Contents

1. Introduction

The project “Generative AI: Navigating Short-Term Skepticism and Long-Term Promise” will address about the challenges and opportunities presented by generative AI in cybersecurity. The main objective of project to balance the current skepticism surrounding GenAI and its long-term potential mainly in the context of cybersecurity applications. The primary goal of the project is to develop a comprehensive strategy for security leaders to effectively integrate GenAI into their cybersecurity practices along with managing associated risks and ethical issues. This project addresses about the gap between the current skepticism and about GenAI’s immediate impact and its promising long-term potential in enhancing cybersecurity measures. In this we are going to discuss about the problems of lack of clear guidance for security leaders on how to navigate the rapid growth of GenAI in cybersecurity and how could we get benefited from GenAI in long runs. Many organizations are struggling to balance the potential benefits of GenAI such as increased productivity and reducing skills gaps. Apart from that we must be careful about the arising disadvantages, risks and ethical considerations because of GenAI.

Here we have discussed about the problems arises on effectively integrating GenAI due to short-term skepticism and uncertainty about long-term benefits.

1. Lack of clear implementation strategies: organizations or stakeholders are struggled to create a comprehensive strategy for implementing GenAI into their cybersecurity framework. This includes where and how to implement, prioritization area and how to measure the success.
2. Ethical and security concerns: There are significant areas that we may need to be careful about while implementing GenAI including issues related to data privacy, biasness in decision-making and security risks.
3. Skills Gap and Workforce Adaptation: Many organizations and business area are lack of hand-on experience on implementing GenAI.
4. Risk Assessment and Mitigation: there’s a lack of clear understanding of risk associated with GenAi in cybersecurity especially implementation and potential vulnerabilities. Thay need skilled and trained people to face the considering risks.
5. Regulatory compliance: Many organizations are uncertain about make sure the use of GenAI in cybersecurity because of current and future regulatory requirements and there is rapid nature of changing in feature of GenAI.

6. Hard to convince stakeholders: Many organizations are facing challenges in convincing stakeholders and other team members about the long-term value of AI in cybersecurity, particularly looking at the current skepticism and associated risk.

1.1 Project Aim

The aim of the project is to provide security leaders with the appropriate knowledge, approaches and strategies to navigate the rapid evolution of Generative AI in cybersecurity, addressing both short-term Skepticism and long-term promises. This includes preparing framework of proactive collaboration with the business stakeholders through ethical guidelines for the safe and secure uses of Gen AI that align with the organization's goals. This project also aims to provide a proper framework for the ethical implementation of GenAI such as ChatGPT and Gemini, conduct thorough risk assessments to identify potential threats and vulnerabilities in cyber security and other sectors and develop robust mitigation strategies. Additionally, it seeks to encourage organizations to provide collaborative training programs for both technical and non-technical staff for the aim of being familiar with the use of GenAI, promoting a culture of teamwork among cybersecurity experts, data scientists, and business stakeholders. Ultimately, this project includes the development of GenAI tools and deploy into the websites hosting in AWS web server. The main aim of deploying AI into the website is to detect the threats and vulnerabilities in the cybersecurity of the website.

1.2 Problem from a Business Perspective

Generative AI such as ChatGPT and Gemini have been playing the important role to transform the industries by democratizing access to skills and enhancing productivity. Many businesses are integrating these technologies to drive innovation, efficiency improvement and filling skills gaps. However, rapid implementations of these technologies into business have raised several challenges

1. Intellectual Property Concerns: "The use of generative AI in creative industries raises significant legal issues such as copyright infringement and the ownership of AI-generated works" (Lucchi, 2023). Companies must navigate these complexities to avoid legal issues and ensure compliance with intellectual property laws. For example: the ChatGPT is AI based system that generates results according to training data into the system and user input, so it is difficult to identify the specific authors.

2. Economic Disruption: “Although with the help of generative AI it is possible to create new jobs and improve the quality of life, the negative effect might concern the loss of jobs and the transformation of traditional industries” (Chui et al., 2023). Generative Ai could identify and prioritize the sales leads by evaluating customers profiles and their priority and customers preferences of choosing products using AI so the organization can make a business plan by evaluating such data. Employers need to weigh this positive angle of AI and its capacity to spur innovation with the fact that AI is likely to radically alter the jobs that are available. When AI work together with the workers it may increase the productivity and accelerate the working period, many others may need to left the job at the same time.

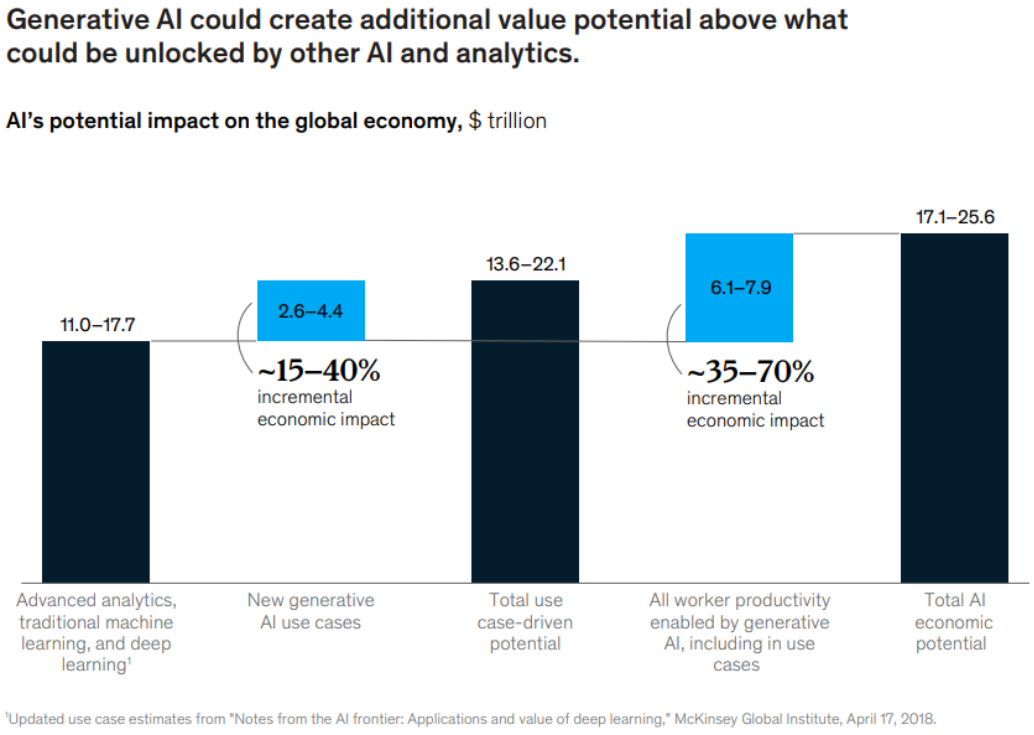


Figure-1 Source: article (McKinsey & Company, 2023)

3. Managing Expectations: In the case of generative AI, more particularly, there is a need to manage expectations in the short-term while keeping an eye on their long-term possibilities. This is because to avoid disillusionment and attain sustainable adoption it is necessary to promote the strategies that are achievable in the business arena. “For example: the use ChatGPT in business and similar LLM has becomes essential tools in business management used in customer service, data analysis, marketing and many other fields” (Rane, 2023). However, they must face the complexity of ethical

dilemmas to technical limitations. Balancing innovation and responsibility is crucial to maximize the benefits of such AI in business. Other challenges of Gen AI similar to ChatGPT and Gemini in the business managements are limitation in contextual understanding, lack of industry base knowledge, ethical considerations and bias, data privacy, Integrity, security vulnerability etc.

1.3 Problem from a Technical Perspective

From a technical point of view, the implementation of generative AI in cybersecurity and other fields involves several challenges:

1. Security Threats and Vulnerabilities: The use of generative AI in cybersecurity creates new risks and attacks on security field. It is crucial for organizations to evaluate the risks likely to be faced to come up with proper measures of preventing the risks that are likely to affect the systems and data.
2. Ethical Implementation: For this reason, it is important that rules of ethics are set for the use of generative AI. This entails putting in place structures that will govern the use of AI tools in a responsible, safe, and value system compliant manner.
3. Data Quality and Bias: The effectiveness of generative AI models depends heavily on the quality and diversity of the data used for training GenAI. Addressing issues related to data bias and ensuring the ethical sourcing of data are critical for the reliable deployment of AI systems.
4. Technical Expertise and Collaboration: Successful implementation of generative AI requires collaboration between cybersecurity experts, data scientists, and business stakeholders. Developing training programs to enhance the skills of both technical and non-technical staff is essential for fostering a collaborative environment and ensuring the secure use of AI technologies.

By addressing these business and technical challenges, organizations can harness the potential of generative AI while mitigating risks and ensuring ethical and secure implementation.

1.4 Other Cybersecurity Related Issues

In this project we will develop AI and integrate into website and the website will be hosting on AWS web server. The main purpose to integrate AI into website is for the security purpose. The AI will be designed to detect vulnerabilities of website and make it secure. During this process several cybersecurity issues may arise. Some of them are listed below:

1. Data protection and privacy: Implementing and processing AI tool chatbots or nay other AI, it may store and access sensitive user data. Poor encryption and access control may result in unauthorized access and breaches.
2. Web application security: misconfiguration in AWS, such as inadequate firewall setting and publicly access resources can lead the application to attack.
3. Social engineering attacks: "Social engineering attacks means the manipulation of individuals informing actions for revealing confidential information from individuals" (Gupta et al., 2023). In the context of cybersecurity this can be attempted to for unauthorized access, sharing sensitive data such as password or credit card number.
4. Phishing attack: "Phishing attack is type of cybercrime where attackers pose trustworthy entities to extract information from victim" (Gupta et al., 2023). Advanced AI, such as ChatGPT, Gemini etc can potentially be exploited by these attackers to make their fishing attempts more effective and harder to detect.

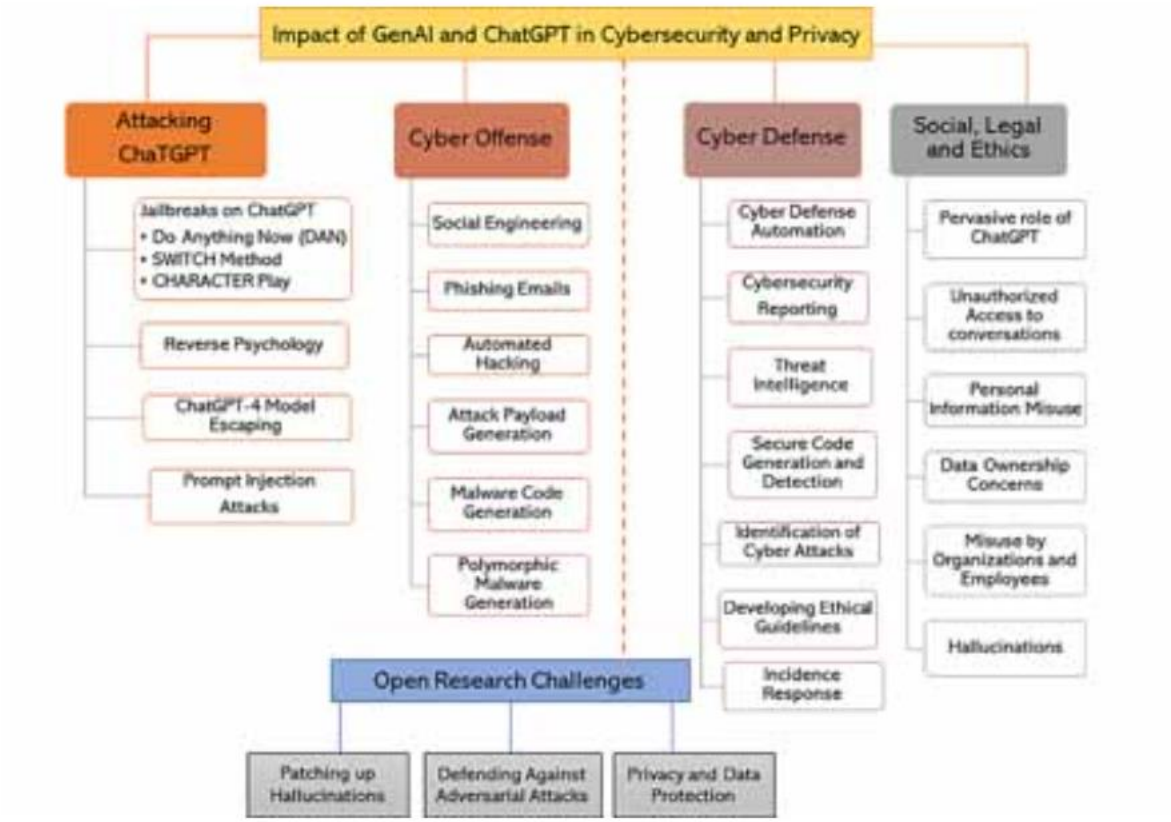


Figure-2: Roadmap of GenAI in cybersecurity and privacy (Gupta et al., 2023)

To address these cybersecurity concerns we can consider the following ideas:

To address the cybersecurity related issues, we can use AI to detect the vulnerability. Apart from that AI can help fix other related issues:

1. Password protection: “By training GenAI over large password datasets, we can make algorithms of identifying structure and patterns in commonly used passwords, so GenAI assists in passwords security assessments” (Dhoni and Kumar, 2023). Gen AI can also be used in users’ behaviour patterns such as password usages, login patterns, password changes etc. that may help Ai to detect unusual or abnormal behaviour that may point to unauthorized access to the system thus helps in potential security violations.
2. Vulnerability scanning and filtering: “The GenAI can be trained to compromise false positives which allow them to learn to generate filters that distinguish actual vulnerabilities from benign and hence reduce false positives in vulnerability scanning” (Dhoni and Kumar, 2023). Gen AI can also be used to effectively scan various programming languages for vulnerability and detects insecure code.
3. Threats hunting queries: LLMs like ChatGPT, Gemini etc can be used to create threat-hunting queries like queries for malware research and detection tools, that enhance response and detection time. They can make a job easy to security group by automatic analysing security incidents.
4. Amazon location service: implement amazon location service to track users’ location securely that helps chatbots to provide location-based responses. Identifying users’ location helps to distinguish the actual intention of the users.
5. AWS CloudFront: implement AWS CloudFront to securely handle the users’ requests. CloudFront assist to pass the original client IP address to the application we are using.
6. Other security protection layers: we can use other security protection layers such as VPC, internet getaway, load balancer, TSL, AWS WAF etc.

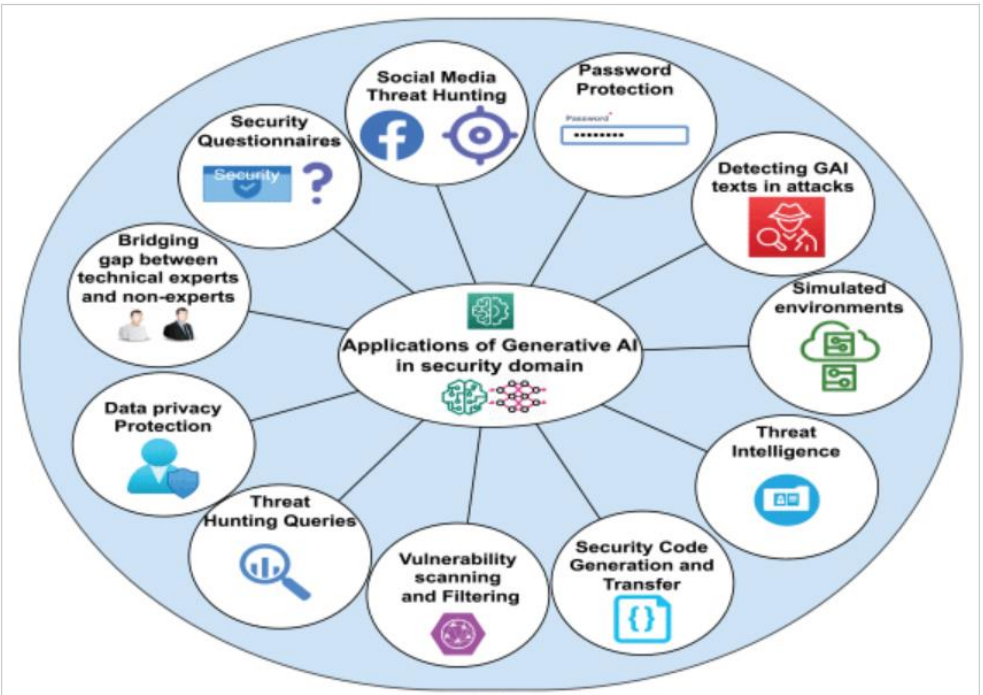


Figure-3: Application of AI in cybersecurity (Dhoni and Kumar, 2023)

1.5 To Solve these Problems, the Project will Propose

1. Recommendations for active collaboration between security leaders and business stakeholders for ethical and secure use of GenAI.
2. Create framework for GenAI implementation in cybersecurity.
3. Risk assessment and mitigation plan for applicating GenAI in cybersecurity.
4. Designing framework of collaborating training and awareness programs to both technical and non-technical staff.

This project is mainly conducted for an industry partner such as large organizations and business sectors, whoever is planning to use GenAI into their Cybersecurity practises. The security leaders, IT professionals and whole business stakeholders might be benefited through this project report. Here we have discussed about some benefits:

1. Clear guidelines for implementing GenAI into cybersecurity practises.
2. Risk management idea and idea about ethical AI implementation.
3. Ideas on how to convince security team, IT team and business team to implement Gen AI.
4. Awareness, understanding and important of GenAI across the organization cybersecurity.

5. Long-term benefits of GenAI in cybersecurity.
6. Developing specific GenAI tools or applications.

This project will not include particularly following ideas.

1. Implementing the recommendations directly within the real-world organization.
2. Provide legal advice on AI regulations.
3. Address GenAI implementing apart from cybersecurity contexts.

This project aims to provide a balanced perspective on GenAI in cybersecurity, acknowledging the current skepticism while preparing organizations to harness its long-term potential. By focusing on ethical implementation, risk management, and collaborative approaches, the project will equip security leaders with the tools and knowledge needed to navigate the evolving landscape of GenAI in cybersecurity

1.6 Literature Review

Generative AI has emerged as a powerful technology providing potential benefits to the various industries including cyber-security. The advancement of Large Language Model (LLM) such as ChatGPT and Gemini have marked significant milestones. However, it has raised significant concerns regarding the ethical implications and potential misuses of GenAI in cybersecurity. Security leaders and organizations have faced both benefits and challenges for implementations of GenAI. The benefits could have been increased productivity, skills enhancement, operational efficiency and cost and at the same time raised concerns about data security, vulnerability and ethical considerations. Establishing guidelines for proactive collaboration with the business stakeholders is crucial for the successful implementation of GenAI in cybersecurity to ensure safe and secure standards. A detailed framework which addresses bias mitigation, transparency, fairness and accountability for ethical AI implementations is essential to guide security leaders are making responsible decisions. A thorough risk assessment is critical to identify potential security threats and vulnerabilities with proper mitigation strategies. As a part of the project, we will develop and implement how GenAI works.

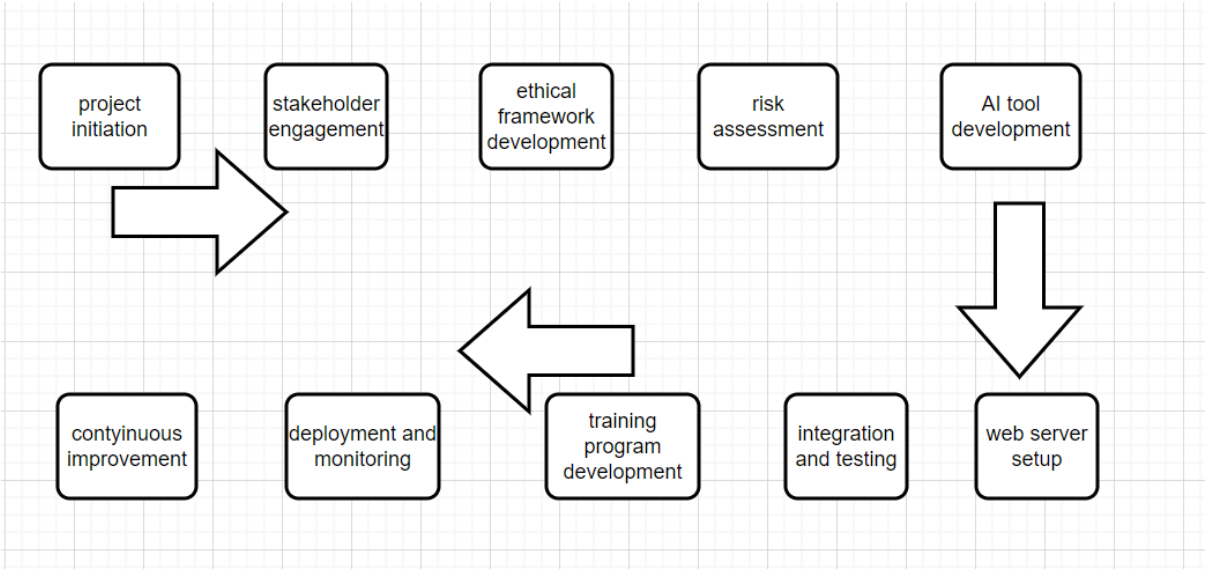


Figure-4: flowchart of whole process

1.7 Agile Methodology for Project Management

For the project "Generative AI: In “Managing for the Short Term while Investing for the Long Term,” the Agile framework will be specific to responding to the needs and volatility of creating generative AI. By using this strategy, it will be possible to promote the cyclical growth of distinct project phases, involve all the stakeholders actively and respond to changes effectively, thus guaranteeing successful project outcomes.

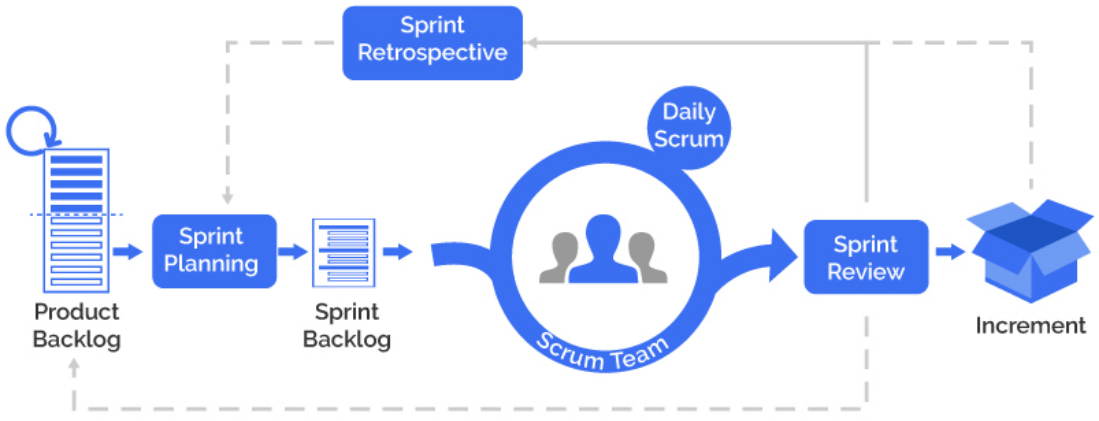


Figure-5 Agile Methodology

- **Product Backlog**: One consolidated list of the project tasks and the specifications.

- **Sprint Planning**: Gives emphasis where work is to be focused.
- **Sprint**: It means that there is a cycle of development that repeats several times with the constant participation of stakeholders.
- **Daily Standup**: Synch meetings for status reports.
- **Sprint Review & Retrospective**: Checking and possible future modifications.
- **Increment**: Any artefacts that are produced during a sprint but are ready to be reviewed or released during the subsequent sprint.

Key Components:

1. **Product Backlog**:
    - **Description**: A schedule of activities and products that are most wanted in order to attain the successful end of the work. For this project, the backlog will include items such as:
        1. Industry analysis of generative AI.
        2. Identification and engagement with key business stakeholders.
        3. Development of an Ethical AI Implementation Framework.
        4. Design and testing of a generative AI chatbot.
        5. Risk assessment and mitigation strategies.
        6. Collaborative training and awareness programs.
2. **Sprint Planning**: In sprint planning meetings the development team will choose some of those activities from the product backlog as what to work on in the sprint. This is a process of planning which entails the assessment of overall working effort and assigning of specific targets towards a particular sprint.
3. **Example**: The players in one sprint may include carrying out industry analysis and initial interviews with the stakeholders.
4. **Sprints**:

- **Description**: Sprint will be time boxed iterations, usually ranging from 2 to 4 weeks. Throughout each sprint, the team with assign itself to the selected tasks with the goal of coming up with a potentially shippable version of the particular project.

5. **Example**: A sprint could be creating the first version of the Ethical AI Implementation Framework and then, piloting it with the stakeholders.
6. **Daily Standup Meetings**:
    - **Description**: Brief conferences during which the individual members report their working progress, issues, and their working plan for the day.
7. **Purpose**: As a result of that to ensure an ideal setting where matters related to curriculum are addressed, there is a need to have and assure an absolute collaboration between the concerned parties to allow for a better ways of handling items of concern within the shortest time possible.
8. **Sprint Review**:
    - **Description:** Each sprint has to be concluded with a presentation of the work done to the stakeholders. It also permits an evaluation to be implemented to ensure that the project formulated corresponds to stakeholder expectations and needs.
    - **Example**: Using the feedback from the stakeholders on the functioning and the ease of the chatbot when presenting the prototype.
9. **Sprint Retrospective**:
    - **Description**: A meeting held at the end of every sprint so that individuals and teams can identify strengths, weaknesses, and opportunities for improvement, for the forthcoming sprint.
    - **Purpose**: To foster a culture of continuous improvement within the team.
10. **Increment**:
    - **Description**: The result of each sprint, which could be a functional component of the project ready for demonstration or further development.

- **Example**: An increment might be a functioning chatbot capable of answering basic queries, ready for further refinement.

1.8 Technical Requirements

This project has two important milestones:

Milestone 1: data collection and project descriptions

In terms of technical requirements for this project, we should have deep understanding of large language models (LLM) like ChatGPT and Gemini and their applications in cybersecurity. A strong foundation in cybersecurity principles and threat detections is essential for developing risk management strategies and mitigations plans. Similarly, knowledge of ethical AI implementations and governance framework is necessary for ethical AI implementation frameworks. For the security component we will deploy VPC attached with internet gateway, load balancer to distribute incoming traffic across AWS web server and we use TLS, DDoS or AWS WAF for the security purposes. Then, we will do the security penetration testing on a system and see the results on cyber security concerns.

Milestone 2: AI tool development and deployment

In this milestone we are going to develop AI and deploy into websites. The technical requirement for this milestone is the subscriptions of Microsoft Azure / AWS web server. For this milestone we have deployed WordPress into AWS web server to run website and completed the task of website design and development. Other technical requirements are to develop a AI and integrate into website. The technical requirements for this task are strong knowledge of programming language python and running code into VS code. Proper arrangements of LLM into the AI system is essential to show the proper functioning of AI like other GenAI ChatGPT and Gemini. Apart from that the requirements are the security tools for the website so that the AI tool should be able to detect and report vulnerabilities and threats and restrict such unidentical activities. For that purpose, we can we web applications firewall, transport layer security etc.

1.9 Specification of Requirements

Developing chatbots used in customer service is a part of our project. This will help us to understand practically the short-term limitations of generative AI followed by long-term benefits.

Functional Requirements

The Generative AI ChatGPT, Gemini, Chatbot must be able to understand and response customer queries in a natural language in a relevant topic to the specific industry. It should be well equipped with the latest information base to provide correct information and resolutions. The system must be able to handle multiple concurrent conversations and seamlessly escalate complex issues to human agents when necessary. Integration with existing customer relationship management (CRM) systems is essential for maintaining context and customer history. Also, such AI should be able to detect the vulnerability issues regarding the cybersecurity.

Usability Requirements

The Generative AI should be intuitive and user-friendly, accessible across various devices and platforms supporting multiple languages and offering various options like text, voice, and visual interactions. They should be able to understand context of the conversation and maintain conversation flow, that help to provide clear and concise responses. There should always be implemented a feedback mechanism to continuously improvement of GenAI performance based on user interactions.

Reliability Requirements

The GenAI such as ChatGPT, Gemini, Chatbot and any other AI tools must operate with high availability and should be able to provide feedback as per the customers' expectations. It should be able to handle sudden increase in user traffic without significant degradation in performance. The system must have robust error handling and fallback mechanisms to ensure continuous operation even when faced with unexpected inputs or system failures. The data must be backed up frequently and have proper disaster recovery plans so that there is no data loss or the time it takes to recover from such a loss.

Performance Requirements

The GenAI should be an always-on and available to answer questions as quickly as possible, ideally within 2 seconds of the question being asked. It must support concurrent users and should be able to increase their performance with the increasing in number of users. The system should be optimized for low latency and efficient resource utilization, ensuring smooth performance even during peak hours.

Security Requirements

Security of the customer data is important, and any form of customer data leak must be prohibited in the GenAI, this includes all forms of communications should be encrypted, secure authentication methods should be used in addition to the compliance with data protection laws. The GenAI must be privacy preserving, meaning that it should only gather and store necessary information and there should be frequent security checks and penetration tests to find and fix potential vulnerabilities. Additionally, the system should have measures against misuse of the AI, for example, such as generating harmful or biased content Selection of Network and Security Technologies

The selection of network and security technologies to address the requirements of GenAI system implementations is essential. The identification, selection, and justification of appropriate network and security technologies to solve the problem of implementing Generative AI while addressing cybersecurity concerns should focus on the following key areas:

Secure AI Platform

Select a secure AI platform such as robust and secure cloud-based infrastructure should be implemented specifically designed for developing and deploying GenAI. Amazon web server and Microsoft azure are recommended for their scalability, reliability and comprehensive secure features. The platform should include robust authentication and access control mechanisms, end-to-end encryptions and secure APIs for integration with existing system.

TLS (Transport Layer Security) for GenAI

"Transport layer security plays an important role to ensure secure communication because of end-to-end encryptions" (Granjal, 2013). The design of TLS security model is helpful for pre-trained large language model such as ChatGPT, Gemini. TLS helps to maintain the privacy of the information exchanged between device and the GenAI server.

OAuth 2.0 and OpenID Connect

"OAuth 2.0 and OpenID Connect can be used to manage authentication and access control" (Thorgersen and Silva, 2021). OpenID Connect can be used to authenticate users securely and obtain information about the user, such as their identity and profile information. This protection also helps to protect against DDoS (distributed denial-of-service) attacks and ensure high availability, employ cloud native services like AWS shield or Azure DDoS protection.

Web Application Firewall (WAF)

“Web application firewall can help protect GenAI applications from common web attacks such as SQL injection, cross-site scripting etc” (Gupta et al., 2023). WAF inspect incoming traffic and filtered out malicious requests from reaching to the GenAI applications. For example: WAF detect and mitigate malicious bot traffic targeting to GenAI applications. Overall, by deploying WAF in front of GenAI applications, organizations can enhance security postures for their AI applications, protect sensitive data and maintain availability and performance of GenAI services.

1.10 Logical/Physical Network Design

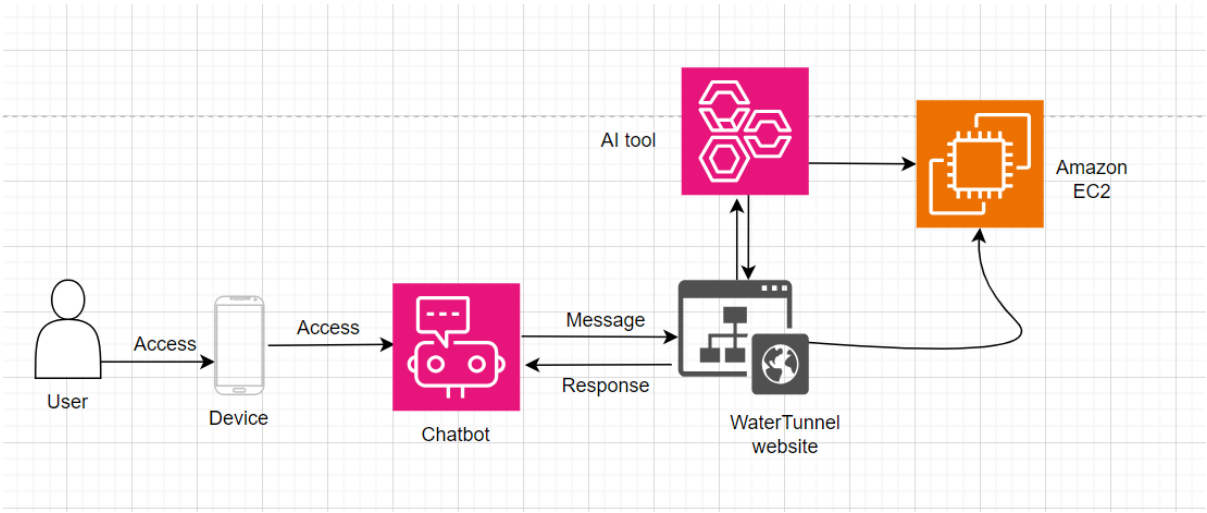


Figure-6: Network Design

In the above figure, AI is associated to the website detects the threats and vulnerabilities in the cybersecurity. The website and AI tool both are hosted on EC2 instances under AWS webserver. In addition, while implementing chatbot a user sends an enquiry message through a website, the chatbot will respond to the message from customers. The aim main of the implementing AI tool in this project is to detect cybersecurity vulnerabilities. The AI tool is designed to detect vulnerabilities of the website. The user or security leasers input the URL of the website, AI tool generate vulnerability report that assists and alerts security leaders to make safe secure the company systems. This report will discuss the security proposals for the entire company system just below. All these systems and data is stored and monitored by AWS cloud services.

1.11 Design of Network / Security Architecture

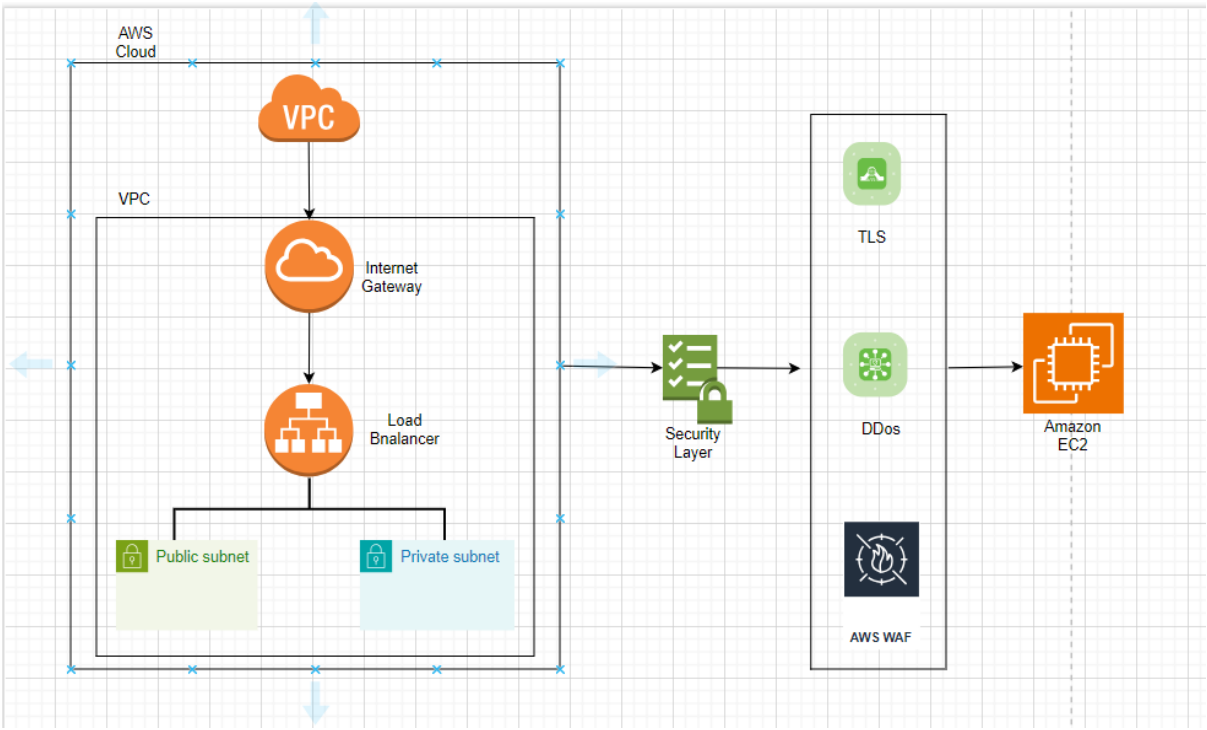


Figure-7: Network/security design

Network Architecture/Protocols/Algorithms

While deploying AI tool and websites on AWS cloud servers, the network architecture plays a crucial role for ensuring secure and efficient communication between clients and servers. Here is a description of the key components of the network architecture:

1. AWS Cloud: The entire system is deployed on the AWS cloud infrastructure, which provides scalability, reliability, and a wide range of services to support the deployment of applications and services.

2. Virtual Private Cloud (VPC): A VPC is created to logically isolate the network resources of the system within the AWS cloud. It allows to define a virtual network environment with its IP address range, subnets, route tables, and network gateways (Patibandla, K.R., 2024).

3. Internet Gateway: An Internet Gateway is attached to the VPC to enable communication between instances in the VPC and the internet. It allows inbound and outbound traffic to and from the internet, facilitating the user's interaction on the website.

4. Load Balancer: A load balancer is used to distribute incoming traffic across multiple AWS web servers hosting the website and AI tool. This helps in load distribution, improves availability, and provides fault tolerance by ensuring that no single server is overwhelmed with traffic.

5. Public Key and Private Key:

- Key Infrastructure (PKI) is implemented in the network architecture to secure communications and authenticate entities. “Public keys are used for encryption and verification, while private keys are used for decryption and signing” (Patibandla, K.R., 2024).
- Public and private key pairs are used for establishing secure connections, such as SSL/TLS connections, between clients and servers. The public key is shared openly, while the private key is kept secure and known only to the server.
- Public and private keys play a crucial role in ensuring secure communication channels and protecting sensitive data transmitted over the network.

Security Architecture/Protocols/Algorithms

In the deployment of websites and AI tool hosted on AWS cloud servers, a comprehensive security architecture is essential to protect the system from various threats and vulnerabilities. Here is a description of the key components of the security architecture:

1. Transport Layer Security (TLS):

    - TLS is implemented to secure communication between clients (e.g. website visitors) and servers (AWS web servers hosting the website).
    - “TLS encrypts data in transit, ensuring that sensitive information exchanged between clients and servers is protected from eavesdropping and tampering” McKay, K. and Cooper, D. (2017, p. 01).
    - By using TLS, the security architecture ensures the confidentiality and integrity of data transmitted over the network.

2. Distributed Denial of Service (DDoS) Protection:

    - DDoS protection mechanisms are implemented to defend against DDoS attacks that aim to disrupt the availability of the customers interaction to the website.

- Utilizing AWS Shield, a managed DDoS protection service, helps safeguard the system from volumetric and application layer DDoS attacks by detecting and mitigating malicious traffic.

3. AWS Web Application Firewall (WAF):

    - AWS WAF is deployed to protect the web application (website and AI tool) from common web exploits and vulnerabilities (Lakhno et al., 2022).
    - It allows security rules to be defined to filter and monitor incoming web traffic, blocking malicious requests before they reach the web servers.
    - AWS WAF works in conjunction with the load balancer to provide an additional layer of defence against SQL injection, cross-site scripting (XSS), and other security threats.

with all those above-mentioned network and security architecture, we can enhance security of the website and AI tools. Even tools will be able to notice unethical cyber threats analysing the intension of the users. Including above mentioned security protocols and algorithms and implementing network access control lists within the VPC we can inbound and outbound traffic at the traffic at the subnet level. Configuring ACL we can restrict traffic based on IP addresses and ranges that will help to effectively restrict access from unauthorised access.

1.12 Network and Security Policies

Password Policies

While implementing Gen AI tool, password policy plays an important role to mitigate cybersecurity risks. When assessing cybersecurity issues, the following password policies would be the best:

1. Increase password length and complexity
2. Enforce password uniqueness
3. Implement multi-factor authentication
4. Secure password storage and transmission
5. Implement account lockout policies
6. Regular password audits

Disaster Recovery and Business Continuity Plan

Comprehensive Disaster Recovery and business Continuity Plan for GenAI Application and associated systems.

Objectives

Ensure Data Integrity: Protect and recover data to ensure business continuity.

Minimize Downtime: Restore the GenAI application and associated services as quickly as possible.

Maintain Communication: Ensure effective communication during and after a disaster.

Compliance: Meet regulatory requirements and maintain cybersecurity standards.

Scope

The Disaster Recovery Plan covers:

- GenAI Application
- Supporting IT infrastructure (servers, APIs, databases)
- Data management (customer data, proprietary information)
- Communication processes and systems.
- Business operations across various potential disruption scenarios.

Disaster Scenarios

- Natural Disasters: Earthquakes, floods, hurricanes
- Cyber-Attacks: All possible cyber-Attacks including Ransomware, data breaches, DDoS and so on.
- Hardware Failures: Server crashes, network failures
- Human Errors: Accidental deletion, misconfigurations
- Software Failures: Bugs, performance issues

Roles and Responsibilities

- Disaster Recovery Team (DRT): A dedicated team responsible for executing the DRP.
- Disaster Recovery Coordinator: Oversees the entire recovery process.
- IT Recovery Lead: Manages IT systems and infrastructure recovery.
- Data Recovery Specialist: Focuses on data integrity and restoration.
- Communication Manager: Handles internal and external communications.

Preparation and Prevention

- Regular Backups: Schedule daily backups of GenAI application data, including customer data and training datasets. Store backups in secure, off-site locations.

- Redundant Systems: Implement failover systems and load balancing to ensure high availability.
- Update and Patch Management: Regularly update and patch GenAI software and underlying systems.
- Employee Training: Train staff on disaster recovery procedures and their roles in a disaster scenario.
- Documentation: Maintain up-to-date documentation of all systems, configurations, and recovery procedures.

Recovery Procedures

- Incident Detection and Notification:
    - Detect and assess the severity of the incident.
    - Notify the Disaster Recovery Team and key stakeholders.
- Assessment and Strategy:
    - Evaluate the impact on the GenAI application and related systems.
    - Develop a recovery strategy based on the type and extent of the disaster.
- Execution of Recovery Procedures:
    - Data Recovery:
        - Restore data from backups.
        - Verify data integrity and completeness.
    - System Recovery:
        - Restore IT systems and infrastructure from backups or redundant systems.
        - Address any hardware or software issues.
    - Application Recovery:
        - Reinstall or repair the GenAI application.
        - Ensure all configurations and integrations are intact.
    - Testing:
        - Conduct thorough testing to verify that systems and applications are fully functional.
- Communication:
    - Communicate with stakeholders, including customers and internal teams, about the status of the recovery process.

- Establish multiple communication channels (e.g., email, SMS, social media, and phone hotlines) to ensure that messages can be delivered even if one channel is unavailable.
- Prepare pre-approved messaging templates for various scenarios, ensuring that accurate and consistent information is delivered quickly.
- Designate a spokesperson and develop a media management plan to handle press inquiries and public statements.
- Provide regular updates and estimated timelines for recovery.

- Post-Recovery:
- Review and document the recovery process and any issues encountered.
- Conduct a post-incident analysis to identify lessons learned and areas for improvement.
- Update the Disaster Recovery Plan based on the findings and new insights.

Testing and Maintenance

- Regular Testing: Conduct regular drills and tests of the disaster recovery procedures to ensure readiness.
- Plan Updates: Review and update the DRP annually or following significant changes to systems or infrastructure.
- Feedback Loop: Incorporate feedback from tests and real incidents to refine the DRP.

Business Continuity Strategies

- Disaster Recovery as a Service (DRAAS): Without the requirement for specialized infrastructure, DRAAS offers a managed disaster recovery solution that guarantees an organized, safe, and effective recovery.
- Zero Trust Data Security: Zero trust security frameworks guarantee rigorous access control and ongoing verification, irrespective of the user's location or network, in order to safeguard data. Install identity and authorization management (IAM) solutions to make sure that only authorized users have access to sensitive information and systems (Snedaker, 2013). Use behavioural analytics and continuous monitoring to identify and address irregularities quickly, stopping illegal access and data breaches.
- Operational Continuity: To enable employees to continue working from off-site locations, establish remote work protocols, such as secure VPN access, collaboration tools, and remote desktop solutions. Determine and set up backup locations for operations (such as data centres or backup offices) so they are ready to go live in case the primary location is compromised. reassigning employees to other positions or

locations as necessary to maintain vital business operations. Create a customer communication strategy to update clients on the state of operations, including any anticipated disruptions and the timeframe for recovery (Snedaker, 2013).

- IT Infrastructure Resilience: To keep services, data, and apps accessible in the event of an interruption, an IT infrastructure must be resilient. To guarantee that vital IT services are still accessible if core systems fail, implement redundant servers, network components, and storage systems. By using load balancing, you may prevent overloading and guarantee high availability by dividing traffic among several servers. Utilize cloud or hybrid cloud solutions to offer resilient, scalable infrastructure that can change to meet evolving business requirements. To avoid malfunctions and guarantee that systems are operating at peak efficiency, do routine maintenance and updates on IT infrastructure.

Compliance and Regulatory Requirements

Verify the Disaster Recovery Plan conforms with all applicable cybersecurity and data protection laws, such as the General Data Protection Regulation, and requirements unique to the industry.

2. System Overview

This project report will deliver following recommendations through this project:

2.1 Industry Analysis for GenAI

The word cybersecurity means a sets of technologies, processes and practices to protect and defend data, network, software and other devices from being attacked, damaged and unauthorized access. In these days the issues related to cybersecurity is very vast, leading to significant growth to cyberattacks for different purposes results that the important of securing IoT. “National institute of Standard and Technologies also encourage to the use more proactive and adaptive approaches towards real-time assessments, continuous monitoring and data analysis to identify, protect against, detect to, and catalogue cyberattacks to prevent future security incidents” (Kaur et al., 2023).

For all those cybersecurity issues, integrating AI tool that can provide analytics and intelligence to protect against cyber-attacks by tracking variety of cyber threats in advance to the problems. For this reason, AI can be used to automate secure tasks and support to the work of human

security teams. Now a days, AI-based cybersecurity tools have been emerged to help security teams to identify risk and improve security. For this purpose, NIST has proposed cybersecurity framework to protect, detect, react and defend against cyber-attack.

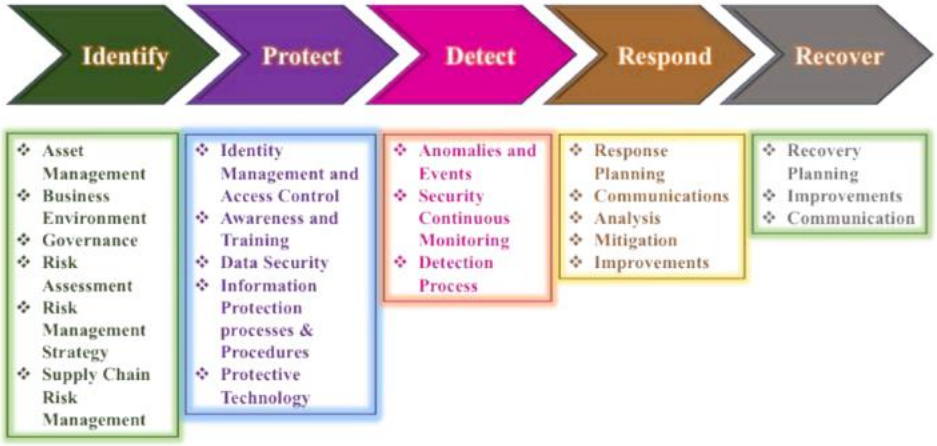


Fig. 1. NIST cybersecurity framework.

Figure-8: NIST cybersecurity framework (Kaur et al., 2023)

The importance of cybersecurity protection is in all industries. For example: Health industry, IT industry, Government website, Automotive industry, banking industry etc where large number of sensitive data are stored, and high number of customers interaction includes. For securing such critical information belongs to individuals it is demanded that cybersecurity protection should be strong. For such security purpose we can integrate AI to provide extra layer for the security.

2.1.1 Automobile Service Industry

For this project we have chosen Automobile services industry "WaterTunnel Car Wash" where we will discuss about the cybersecurity problems in automobile industry solutions and related issues here:

We have seen the significant rise in cybersecurity issues in Automobile service industry in recent years. This report will discuss about the importance of AI in cybersecurity in Automobile service industry (Water Tunnel Car Wash company). Implementing AI in such industry have provided significant benefits to the organizations and has also helped to solve the privacy and security issues.

Cybersecurity related issues and solutions using GenAI tool to Automobile services industry:

1. **Data privacy and security concerns:** Ensuring data security and privacy is one of the biggest issues facing the automobile service industry, especially for businesses like Water Tunnel Car Wash. Customer transactions are involving more and more sensitive personal information exchanged, including names, contact details, payment information, and even vehicle-related data, as the industry gets more digitalized. This data is frequently kept in databases and accessed by a variety of technologies, such as AI-powered chatbots intended to improve customer support. Nevertheless, there are a lot of hazards associated with processing and storing such data.

   Insufficient protection of this sensitive data may result in serious privacy breaches, giving unauthorized parties access to customers' personal information. Such hacks may lead to identity theft, monetary losses, and serious harm to the business's standing. The application of AI to cybersecurity offers a proactive strategy to reduce these risks by continually and instantly monitoring for any irregularities or attempts at illegal access. AI technologies can automatically identify possible security risks, allowing for quick action to stop data breaches.

2. **Network Security and Vulnerabilities**

   Network security is a significant issue in the automobile sector. For several functions, including as customer support, payment processing, and operational logistics, modern automakers mainly rely on networked systems. However, because of network flaws including shoddy software, unprotected Wi-Fi networks, and inadequate encryption, cybercriminals frequently target these networked systems.

   “Ransomware, for example, is a type of hack that can cripple an entire network and prevent access to vital data and services unless a ransom is paid” (Yaqoob et al., 2017). By using machine learning algorithms to find odd patterns of behaviour inside the network, artificial intelligence (AI) can play a key role in locating and addressing these vulnerabilities in the network. By examining historical events and present danger landscapes, AI can also anticipate possible attack vectors, enabling businesses to fortify their network defences prior to an attack.

3. **Secure AI Integration**

   Care must be taken when integrating AI into an automobile company's cybersecurity architecture. Even though artificial intelligence (AI) has many advantages, it also poses new security risks. For example, there is a chance that hostile actors will exploit AI algorithms to get beyond security safeguards. Secure AI integration means Companies

need to take a multi-layered security approach to combat this, which involves ongoing algorithm changes, continuous performance monitoring of AI, and secure AI development methods (Habbal et al., 2024). Because AI technologies offer dynamic threat detection and response capabilities, they may also be used to protect the systems that they are a part of. AI-driven systems, for instance, can automatically modify security procedures in reaction to new threats, guaranteeing that the business's cybersecurity defences continue to be strong even in the face of constantly changing cyberthreats.

4. **Regulatory Compliance and Ethical Considerations.**

   Like many other industries, the automobile service industry is governed by stringent laws pertaining to cybersecurity and data protection. Maintaining customer trust and avoiding legal ramifications necessitates compliance with rules such as the California Consumer Privacy Act (CCPA) in the US and the General Data Protection Regulation (GDPR) in Europe (Bakare et al., 2024). By automating the tracking of data usage, access controls, and permission management, artificial intelligence (AI) can help ensure compliance and lower the risk of regulatory violations.

   Equally crucial is the moral use of AI in cybersecurity thought, businesses need to make sure AI solutions are transparent and protect client privacy when they are built and deployed. Customers must be informed about the collection, storage, and use of their data. Additionally, judgments made by AI-powered systems, like those pertaining to security precautions, must be transparent and equitable.

The Water Tunnel Car Wash company serves as an example of how AI integration into the automobile cybersecurity framework considerable potential has must improve data privacy, network security, and overall operational resilience. Automobile firms preserve regulatory compliance, safeguard sensitive consumer data, and create a more reliable and secure digital environment by using AI to proactively handle cybersecurity concerns. However, to fully reap the benefits of AI while minimizing associated hazards, significant thought must be paid to its ethical and secure application.

2.1.2 Stakeholders' Identification

Stakeholders on Automobile service industry particularly WaterTunnel Car Wash company has individuals, groups, or organizations that have an interest in, are affected by, or can influence

the outcome of a project, initiative, or business activity. They can be internal (within the organization) or external (outside the organization) and their interests, influence, and involvement can vary significantly. The characteristics of stakeholders are

- **Interest:** Stakeholders have a vested interest in the project's outcome, as it can impact them directly or indirectly.
- **Influence:** Stakeholders can affect the project's direction and decisions through their power, resources, or position.
- **Impact:** Stakeholders can be impacted by the project positively or negatively, and their support or opposition can affect the project's success.

2.1.3 Types of Stakeholders

Internal Stakeholders: These are individuals or groups within the organization, such as employees, managers, and shareholders.

External Stakeholders: These are individuals or groups outside the organization, such as customers, suppliers, regulators, and the community.

Stakeholder	Impact	Engagement
Internal Stakeholders		
Employees	- Customer Service Reps: Changes in daily tasks and responsibilities. - Site Managers: Oversee integration, ensure enhanced customer service. - IT Staff: Technical deployment, maintenance, troubleshooting.	- Provide training, gather feedback - Involve in planning/testing, monitor performance, adjust procedures - Regular technical meetings, detailed project plans, ongoing support.

	- Marketing and Sales Team: Use chatbot, phone call, workshops for engagement, Sales promotions and insights.	
Operations Manager	Oversee integration/effectiveness across locations	Regular updates, address operational challenges, optimize processes.
IT security leaders	Tracking cybersecurity related issues	Ensure ethical, secure and safe use of AI
Shareholders	Interested in financial performance and ROI	Provide reports on cost savings, efficiency improvements, customer satisfaction.
External Stakeholders		
Customers	Interact for inquiries, bookings, Service updates, support, FAQs and feedback	Ensure user-friendly interface, provide information, gather feedback
Regulators	Ensure compliance with regulations (data protection, consumer rights)	Ensure legal compliance, conduct audits, maintain data handling transparency

Table-1 Type Stakeholders

2.1.4 Stakeholder matrix (influence/interest)

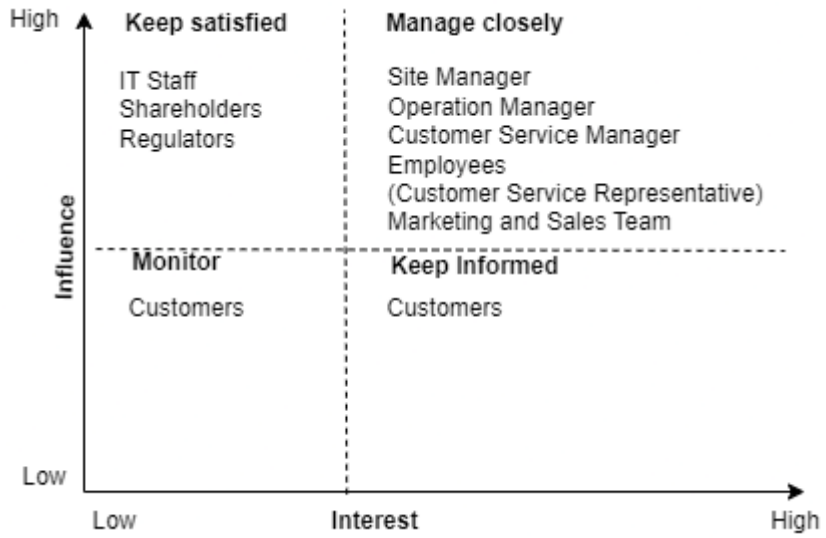


Figure-9 Interest Influence Matrix

The above stakeholder engagement matrix shows that IT staff such as security leaders have high influence but low interest so they should keep satisfied. Managers have high influence and high interest regarding the implementations of chatbots so they should manage closely. Customers have high interest but low influence so they should keep informed. At last, a few other customers have low influence and low interest so they should monitor.

2.1.5 Set of Strategies for Security Leaders

Here's a comprehensive set of guidelines for security leaders to engage in proactive collaboration with business stakeholders for the ethical, safe, and secure use of Generative AI (GenAI) in cybersecurity for the organizations like automobile services (Water Tunnel Car Wash company):

Establish an AI Ethics Committee

Create a special AI Ethics Committee representative from various stakeholders inside the organizations such as departments, cybersecurity leaders, legal advisors, HR, and other business units within the organizational for ethical, safe and secure use of AI inside automotive service industry. This committee will:

- Develop and oversee ethical guidelines for GenAI implementation
- Review and approve AI projects
- Ensure alignment with organizational goals and values

Develop a Clear GenAI Strategy

1. Define Objectives: “Clearly articulate the organization's goals for implementing GenAI in cybersecurity, such as improving threat detection or automating routine tasks etc” (Huang et al., 2024).
2. Prioritize Use Cases: Identify and prioritize specific areas of automobile service industry where GenAI can provide the significant contribution such as vulnerability detection, securing personal information of customers and entire system that supports the organization's cybersecurity efforts.
3. Set Realistic Expectations: Manage stakeholder expectations by clearly communicating both the potential benefits and limitations of GenAI in the short-term and long-term process and understand for all those external and internal stakeholders of the automobile service industry. Potential benefits could be securing company system, pre-identify the security alerts and limitations could be implementations issues and higher cost etc.

Implement a Robust Governance Structure

1. Roles and Responsibilities: Clearly define roles and responsibilities of each stakeholder for AI implementation and oversight across the automobile service industry such as HR has roles and responsibility of proposing plans for GenAI implementation, security leaders must have responsibility of securely implementing GenAI, managers must be aware of educating all staffs and customers for working with the AI.
2. Decision-Making Protocols: Establish clear protocols for AI-assisted decision-making, especially in critical cybersecurity scenarios to make sure the biasness, transparency and accuracy etc.
3. Accountability Measures: Develop accountability frameworks to ensure responsible use of GenAI and address any issues that may arise.

Conduct Comprehensive Risk Assessments

1. Identify Potential Threats: “Regularly assess and update the potential security threats and vulnerabilities associated with GenAI implementation” (Huang et al., 2024). The risk could be false result provided by GenAI due to outdated system, some GenAI may be gateway way for the hackers. We have identified few potential threats in cyber security focusing on automobile service industry, which we will describe in next topic of our report.
2. Develop Mitigation Strategies: Create robust mitigation plans to address identified risks, including data privacy concerns and potential biases in AI models. We have developed a robust mitigation strategy for implementing GenAI into cybersecurity in next topic of our report.
3. Continuous Monitoring: Implement ongoing monitoring systems to detect and respond to emerging risks related to GenAI use.

Ensure Data Privacy and Security

1. Data Management Policies: All the information regarding the company and information belongs to customers are stored in server system, GenAI has responsibility to secure them. Develop strict data management policies for AI training and operation, ensuring compliance with relevant regulations.

2. Encryption and Access Control: Implement strong encryption and access control measures to protect sensitive data stored in the system using in GenAI.
3. Regular Audits: Conduct regular audits of data usage and AI system outputs to ensure compliance with privacy and security standards, sometimes old version of AI systems results negative outcomes.

Foster Transparency and Explainability

1. Documentation: Maintain comprehensive documentation of AI models, including their training data, decision-making processes, and limitations (Balasubramaniam et al., 2022) soi the result provided by GenAI could be dependable.
2. Explainable AI: The security leaders should prioritize the use of explainable AI techniques to ensure that AI-driven decisions in cybersecurity can be understood and justified.
3. Stakeholder Communication: Regularly communicate with stakeholders about the use and impact of GenAI in cybersecurity operations that establish the accountability and transparency among the stakeholders.

Implement Collaborative Training Programs

1. Skills Development: Inside the industry both type of staffs is there, everyone should be aware of implementing GenAI and its working process. So, the security leaders must develop training programs to enhance the AI literacy of both technical and non-technical staff. The training and programs for technical and non-technical staff, we will discuss briefly on the later part of this report.
2. Cross-Functional Collaboration: Encourage collaboration between cybersecurity experts, data scientists, and business stakeholders through joint training sessions and projects inside the automobile service industry.
3. Continuous Learning: Establish a culture of continuous learning to keep pace with rapidly evolving AI technologies and best practices.

Engage in Ethical AI Development

1. Ethical Guidelines: Develop and adhere to clear ethical guidelines for AI development and deployment in cybersecurity contexts. Each industry must have different guidelines for implementing GenAi such as purpose, outcomes and impacts.
2. Bias Mitigation: Implement strategies to identify and mitigate biases in AI models and decision-making processes.
3. Human Oversight: Maintain appropriate human oversight in critical AI-driven cybersecurity processes to ensure ethical considerations are not overlooked.

Establish Feedback Mechanisms

1. User Feedback: Create channels for employees and stakeholders to provide feedback on GenAI implementations and their impacts.
2. Continuous Improvement: Use feedback to continuously refine and improve GenAI systems and their integration into cybersecurity processes.

Ensure Regulatory Compliance

1. Regulatory Monitoring: Stay informed about evolving regulations related to AI use in cybersecurity and ensure compliance.
2. Proactive Adaptation: Develop strategies to quickly adapt GenAI implementations to meet changing regulatory requirements.

These are the guidelines for security leaders that promotes ethical, safe and secure use of GenAI while aligning with organizational goals and values. We have described briefly each topic and sub-topic in later part of our project report.

2.1.6 Organizational Goals

Here are the major goals that we have identified for the automobile service industry regarding GenAI in cybersecurity:

1. Enhance cybersecurity capabilities: The aim of the industry to leverage GenAI to improve threat detection, automate routine tasks, and strengthen overall security posture (Dhoni and Kumar, 2023).
2. Increase operational efficiency: There is a focus on using GenAI to boost productivity and reduce skills gaps in cybersecurity teams that supports cybersecurity teams establishing safe and security system.
3. Manage risks effectively: Another key goal of the industry is to conduct thorough risk assessments of GenAI applications in cybersecurity and develop robust mitigation strategies (Chandrasekaran, A.S., 2024).
4. Foster collaboration: Industry must encourage teamwork between cybersecurity experts, data scientists, and business stakeholders in implementing GenAI.
5. Develop AI literacy: Another goal of the industry is to raise awareness and understanding about GenAI's capabilities, limitations, and best practices across technical and non-technical staff.
6. Navigate regulatory compliance: Industry need to ensure their use of GenAI in cybersecurity meets current and future regulatory requirements.
7. Create clear implementation strategies: Industry must develop comprehensive strategies for integrating GenAI into their cybersecurity frameworks, including prioritization and success metrics.
8. Address ethical and security concerns: There's a focus on tackling issues related to data privacy, bias in decision-making, and potential security risks associated with GenAI.

2.2 Framework for GenAI Implementation

Generative AI (GenAI) is an emerging technology that has been gaining even more momentum in the past few years; parallel to rapid developments on Artificial Intelligence, new capabilities are being leveraged by Blue and Red Teams. Technologies like large language models (LLMs) and generative adversarial networks are part of the broad field of AI, including genAI capabilities such as threat detection/incident response/predictive analysis.

On the other hand, the use of GenAI within cybersecurity brings with it a whole host of ethical implications. That includes biases in AI decision-making, privacy concerns with processing sensitive data and the danger of becoming too dependent on AI systems as well as misinformation arising from or misusing AI technology. While GenAI is being rapidly adopted by organizations to increase their cybersecurity capabilities the need for a strong ethical framework alongside its development, deployment and use becomes imperative.

2.2.1 Comparative Analysis of Best Practices for Ethical Frameworks for GenAI Across Sectors

Across numerous sectors have unique ethical considerations caused by the new content generation capabilities for Generative AI (GenAI), such as privacy, authenticity and intellectual property. Here we look into the best practices for ethical GenAI frameworks of different industries and derive lessons that relevant to cyber security in the section.

Best Practice	Healthcare	Finance	Media & Entertainment	Education	Cybersecurity
Data Privacy & Consent	Strict protocols for patient data	Robust measures for financial data	Content creator rights protection	Student data protection	Sensitive security data handling
Output Authenticity	Medical record generation	Financial report creation	Deepfake prevention	Plagiarism detection	Threat simulation fidelity
Bias Mitigation	Diverse training data for diagnoses	Fair lending algorithms	Inclusive content generation	Equitable educational content	Unbiased threat detection

Transparency	Explainable AI for treatment recommendations	Clear disclosure of AI-generated content	Watermarking of AI-generated media	Visibility into AI-assisted grading	Transparent AI-driven alerts
Human Oversight	Doctor review of AI-generated diagnoses	Human verification of AI financial decisions	Editor review of AI-generated content	Teacher oversight of AI-generated lessons	Analyst verification of AI-detected threats
Intellectual Property	Clear ownership of AI-generated medical insights	Attribution of AI-generated financial models	Rights management for AI-created content	Proper citation of AI-assisted work	Ownership of AI-generated security protocols
Continuous Monitoring	Regular audits of AI system performance	Real-time monitoring of AI trading systems	Ongoing quality checks of AI-generated content	Continuous assessment of AI educational tools	Dynamic evaluation of AI threat detection

Table-2 Comparison of GenAI Ethical framework best practices

Analysis of the comparison reveals several key insights that can be applied to the cybersecurity sector

1. Data Privacy and Consent: The healthcare sector's stringent protocols for handling patient data can inform cybersecurity practices for managing sensitive security information. Implementing HIPAA-like standards for data protection in cybersecurity can enhance trust and compliance.
2. Output Authenticity: The media and entertainment industry's focus on deepfake prevention offers valuable lessons for ensuring the authenticity of AI-generated threat simulations and intelligence reports in cybersecurity.

3. Bias Mitigation: The finance sector's emphasis on fair lending algorithms provides a model for developing unbiased threat detection systems in cybersecurity, ensuring equitable treatment across different network segments and user groups.
4. Transparency: Healthcare's use of explainable AI for treatment recommendations can be adapted to create transparent AI-driven security alerts, enhancing trust and understanding among cybersecurity professionals.
5. Human Oversight: The tiered approach to human verification seen in the finance sector can be applied to cybersecurity, where the level of human involvement correlates with the potential impact of security decisions.
6. Intellectual Property: Clear attribution practices from the education sector can inform the development of protocols for managing ownership and rights of AI-generated security insights and strategies.
7. Continuous Monitoring: Real-time monitoring systems used in finance can be adapted for the dynamic evaluation of AI threat detection in cybersecurity, ensuring ongoing accuracy and reliability.

By incorporating these cross-sector insights, the cybersecurity industry can develop a more comprehensive and robust ethical framework for GenAI implementation.

2.2.2 Ethical Principles

The foundation of ethical AI implementation in cybersecurity rests on six core principles. These principles should guide all aspects of GenAI development, deployment, and use in cybersecurity contexts.

I. Transparency

Transparency is crucial for building trust in AI systems and ensuring accountability. In the context of GenAI in cybersecurity, transparency involves:

- Clear Communication: Organizations must clearly communicate when and how GenAI is being used in their cybersecurity operations. This includes informing relevant stakeholders (e.g., employees, clients, regulators) about the presence of AI systems, their purpose, and their scope of operation.
- Explainable Processes: The decision-making processes of GenAI systems should be explainable to both technical and non-technical stakeholders. While the intricacies of complex AI models may not always be fully interpretable, efforts should be made to provide meaningful explanations of how the AI arrives at its conclusions or decisions.
- Performance Reporting: Regular reporting on the performance and impact of GenAI systems should be conducted. This includes sharing success metrics, limitations, and any significant incidents or errors.

- Open Dialogue: Organizations should foster an open dialogue about their use of GenAI in cybersecurity, encouraging questions and addressing concerns from stakeholders.

Implementation strategies:

- Develop clear communication policies regarding AI use
- Invest in explainable AI techniques and tools
- Establish regular AI performance reporting mechanisms
- Create forums for stakeholder dialogue on AI use in cybersecurity

II. Accountability

Accountability ensures that there are clear lines of responsibility for the actions and decisions of GenAI systems. Key aspects include:

- Clear Ownership: Establish clear ownership and responsibility for GenAI systems within the organization. This includes designating individuals or teams responsible for the development, deployment, and ongoing management of AI systems.
- Auditability: Implement mechanisms for auditing and reviewing GenAI actions. This includes maintaining detailed logs of AI decisions and actions, enabling post-hoc analysis and investigation.
- Human Oversight: Ensure there's appropriate human oversight for critical decisions made by GenAI systems. This may involve implementing human-in-the-loop processes for high-stakes decisions or actions.
- Legal and Regulatory Compliance: Ensure that the use of GenAI in cybersecurity complies with relevant laws and regulations, including data protection and privacy laws.

Implementation strategies:

- Define clear roles and responsibilities for AI system management
- Implement robust logging and auditing systems
- Establish human oversight protocols for critical AI decisions
- Regularly review and update AI systems to ensure legal compliance

III. Fairness and Non-discrimination

GenAI systems in cybersecurity must operate fairly and without discrimination. This principle involves:

- Equitable Treatment: Ensure that GenAI systems provide equitable treatment across different user groups, network segments, and types of cyber threats. The system should not unfairly advantage or disadvantage any particular group.
- Bias Detection and Mitigation: Regularly test GenAI systems for biases in threat detection, incident response, and other cybersecurity functions. Implement measures to detect and mitigate biases in AI models and decision-making processes.
- Diverse Development Teams: Foster diversity in AI development teams to help identify and mitigate potential biases and ensure a range of perspectives are considered in system design and implementation.
- Fair Resource Allocation: Ensure that AI-driven resource allocation in cybersecurity (e.g., allocation of security controls or incident response resources) is based on objective risk assessments rather than biased or discriminatory factors.

Implementation strategies:

- Implement regular bias testing and correction processes
- Establish diversity and inclusion initiatives in AI development teams
- Develop fair resource allocation algorithms and policies
- Conduct regular fairness audits of AI systems

IV. Privacy and Data Protection

Protecting privacy and securing sensitive data is paramount in cybersecurity. For GenAI systems, this principle encompasses:

- Data Minimization: Collect and use only the data necessary for the intended cybersecurity functions. Avoid unnecessary data collection or retention.
- Robust Data Protection: Implement strong security measures to protect data used by GenAI systems, including encryption, access controls, and secure data storage practices.

- Privacy-Preserving Techniques: Utilize privacy-preserving AI techniques, such as federated learning or differential privacy, where appropriate to minimize exposure of sensitive data.
- Compliance with Data Protection Regulations: Ensure that the use of GenAI in cybersecurity complies with relevant data protection regulations (e.g., GDPR, CCPA) and industry standards.
- User Consent and Control: Where applicable, obtain informed consent for data use and provide users with control over their data, including options for data access, correction, and deletion.

Implementation strategies:

- Conduct regular privacy impact assessments
- Implement state-of-the-art data protection measures
- Explore and adopt privacy-preserving AI techniques
- Establish clear data governance policies and procedures
- Develop user-friendly interfaces for data control and consent management

V. Human Oversight and Control

Maintaining appropriate human oversight and control over GenAI systems in cybersecurity is crucial for ethical operation. This principle involves:

- Meaningful Human Involvement: Ensure that humans remain meaningfully involved in critical decision-making processes, especially for high-stakes cybersecurity decisions.
- Ability to Override: Implement mechanisms that allow human operators to override or intervene in AI decisions when necessary.
- Clear Delineation of AI and Human Roles: Clearly define and communicate the respective roles and responsibilities of AI systems and human operators in cybersecurity processes.
- Continuous Monitoring: Establish processes for continuous human monitoring of AI system performance and decision-making.
- Ethical Decision-Making Authority: Ensure that ethical decisions and judgments remain under human control, with AI systems serving as decision support tools rather than autonomous decision-makers in ethically sensitive situations.

Implementation strategies:

- Develop clear protocols for human-AI interaction in cybersecurity processes
- Implement technical safeguards that allow for human intervention in AI systems
- Provide training to human operators on effective oversight of AI systems
- Establish regular review processes to assess the balance of AI automation and human control

VI. Robustness and Security

GenAI systems in cybersecurity must be robust and secure to ensure reliable and trustworthy operation. This principle encompasses:

- Resilience to Attacks: Design GenAI systems to be resilient against adversarial attacks, including attempts to manipulate or deceive the AI.
- Stability and Reliability: Ensure that GenAI systems perform consistently and reliably across various operational conditions and scenarios.
- Secure Development Practices: Implement secure coding practices and rigorous testing procedures in the development of GenAI systems.
- Regular Security Assessments: Conduct regular security assessments and penetration testing of GenAI systems to identify and address vulnerabilities.
- Fail-Safe Mechanisms: Implement fail-safe mechanisms that ensure safe system behavior in case of AI failures or unexpected situations.

Implementation strategies:

- Incorporate adversarial training in AI model development
- Implement robust error handling and fallback mechanisms
- Adopt secure DevOps practices for AI system development
- Establish a regular schedule for security audits and penetration testing
- Develop and test fail-safe protocols for AI systems

2.2.3 Governance Structure

A robust governance structure is essential for ensuring the ethical implementation of GenAI in cybersecurity. This structure provides oversight, defines responsibilities, and establishes processes for ethical decision-making.

I. AI Ethics Committee

The AI Ethics Committee plays a crucial role in overseeing the ethical aspects of GenAI implementation in cybersecurity.

1. Composition: The committee should include a diverse group of experts to ensure comprehensive ethical oversight:

- Cybersecurity experts
- AI and machine learning specialists
- Legal counsel with expertise in technology law and data protection
- Ethics professionals
- Privacy officers
- Representatives from key business units
- External advisors (e.g., academics, industry experts) for independent perspectives

Responsibilities:

- Review and approve GenAI implementations in cybersecurity
- Develop and regularly update ethical guidelines for AI use
- Assess the ethical implications of new AI technologies and use cases
- Address ethical concerns and incidents related to AI use
- Provide guidance on ethical dilemmas in AI implementation
- Review and approve AI-related policies and procedures
- Oversee ethical training programs for staff involved in AI development and use

Operations:

- Regular meetings (e.g., monthly) to discuss ongoing AI projects and ethical considerations
- Ad-hoc meetings to address urgent ethical issues or incidents
- Annual review of the organization's AI ethics posture
- Reporting to senior management and the board on ethical AI matters

II. Roles and Responsibilities

Clear definition of roles and responsibilities is crucial for effective ethical governance of GenAI in cybersecurity.

Key roles:

1. Chief AI Ethics Officer:

- Oversees the overall ethical compliance of GenAI systems
- Chairs the AI Ethics Committee
- Develops and maintains the organization's AI ethics strategy
- Liaises with senior management on AI ethics matters

2. AI Ethics Specialists:
    - Conduct ethical impact assessments for new AI projects
    - Provide guidance on ethical AI development practices
    - Support the AI Ethics Committee in policy development
    - Collaborate with AI developers to implement ethical guidelines
3. Data Protection Officer:
    - Ensures GenAI systems adhere to data privacy regulations
    - Oversees data protection impact assessments for AI projects
    - Advises on privacy-preserving AI techniques
    - Monitors compliance with data protection policies
4. AI Auditor:
    - Conducts regular ethical audits of GenAI systems
    - Assesses AI systems for bias, fairness, and transparency
    - Verifies compliance with ethical AI policies and guidelines
    - Reports audit findings to the AI Ethics Committee
5. AI Development Team Lead:
    - Ensures ethical considerations are integrated into AI development processes
    - Implements technical measures for explainability, fairness, and robustness
    - Collaborates with the AI Ethics Specialists on ethical design choices
6. Cybersecurity Team Lead:
    - Oversees the integration of GenAI into cybersecurity operations
    - Ensures alignment between AI capabilities and cybersecurity objectives
    - Manages human-AI collaboration in cybersecurity processes
7. Legal Counsel:
    - Advises on legal implications of AI use in cybersecurity
    - Ensures AI systems comply with relevant laws and regulations
    - Supports the development of AI-related contracts and agreements

III. Ethical Review and Approval Process

A structured process for ethical review and approval of GenAI projects in cybersecurity is essential to ensure alignment with ethical principles and guidelines.

Process steps:

1. Project Initiation:
    - Project team submits an AI project proposal
    - Initial screening for potential ethical implications
2. Ethical Impact Assessment:
    - AI Ethics Specialists conduct a thorough ethical impact assessment
    - Assessment covers all relevant ethical principles and potential risks
3. Review by AI Ethics Committee:
    - Committee reviews the project proposal and ethical impact assessment
    - Considers alignment with ethical guidelines and potential ethical issues
4. Recommendations and Requirements:
    - Committee provides recommendations for ethical implementation
    - Specifies any requirements or modifications needed for approval
5. Project Modification:
    - Project team addresses the committee's recommendations
    - Makes necessary modifications to ensure ethical compliance
6. Final Approval:
    - AI Ethics Committee reviews the modified project plan
    - Grants approval if ethical requirements are met
7. Implementation and Monitoring:
    - Project proceeds with implementation
    - Ongoing monitoring for ethical compliance
8. Post-Implementation Review:
    - Conduct a review after a specified period (e.g., 6 months)
    - Assess actual ethical impact and effectiveness of mitigation measures

This process should be agile enough to accommodate the fast-paced nature of cybersecurity operations while ensuring thorough ethical consideration.

2.2.4 Risk Assessment

A comprehensive ethical risk assessment is crucial for identifying and mitigating potential ethical issues in the implementation of GenAI in cybersecurity.

I. Identification of Ethical Risks

The first step in ethical risk assessment is to identify potential ethical risks associated with the use of GenAI in cybersecurity. This process should involve a diverse team of stakeholders to ensure a comprehensive view of potential risks. Some key ethical risks to consider include:

1. Privacy Breaches:
    - Unauthorized access to sensitive data processed by GenAI systems
    - Unintended disclosure of personal information through AI outputs
    - Re-identification of anonymized data through AI analysis
2. Bias and Discrimination:
    - Unfair treatment of certain user groups or network segments
    - Biased threat detection leading to false positives or negatives
    - Discriminatory resource allocation in incident response
3. Lack of Transparency:
    - Inability to explain AI decision-making processes
    - Lack of clarity on when and how AI is being used
    - Difficulty in auditing AI actions
4. Overreliance on AI:
    - Erosion of human expertise in cybersecurity
    - Inability to detect AI errors or malfunctions
    - Overdependence on AI for critical security decisions
5. AI Manipulation:
    - Adversarial attacks on AI systems leading to misclassification
    - Exploitation of AI vulnerabilities by malicious actors
    - Use of GenAI to create sophisticated phishing or social engineering attacks
6. Ethical Decision-Making:
    - AI making ethically sensitive decisions without human oversight
    - Conflicts between AI optimization goals and ethical considerations
    - Difficulty in encoding complex ethical principles into AI systems
7. Job Displacement:

- Potential loss of cybersecurity jobs due to AI automation
- Shift in required skills leading to workforce challenges

8. Legal and Regulatory Compliance:
    - Violation of data protection regulations due to AI data processing
    - Non-compliance with industry-specific regulations
    - Liability issues arising from AI-driven decisions

II. Impact and Likelihood Evaluation

Once ethical risks are identified, they should be evaluated based on their potential impact and likelihood of occurrence. This evaluation helps prioritize risks and allocate resources for mitigation efforts.

Evaluation process:

1. Impact Assessment:
    - Evaluate the potential consequences of each identified risk
    - Consider factors such as financial impact, reputational damage, legal implications, and harm to individuals or groups
    - Assign an impact rating (e.g., Low, Medium, High, Critical)
2. Likelihood Assessment:
    - Estimate the probability of each risk occurring
    - Consider factors such as the complexity of the AI system, existing controls, and historical data
    - Assign a likelihood rating (e.g., Rare, Unlikely, Possible, Likely, Almost Certain)
3. Risk Matrix:
    - Plot each risk on a risk matrix based on its impact and likelihood
    - Use this visual representation to prioritize risks
4. Risk Prioritization:
    - Categorize risks as Low, Medium, High, or Critical based on their position in the risk matrix
    - Focus immediate attention on high and critical risks

Example Risk Matrix:

Critical	M	H	C	C	C
High	M	M	H	H	C
Medium	L	M	M	H	H
Low	L	L	M	M	H

Table-4 Example Risk Matrix

L = Low Risk, M = Medium Risk, H = High Risk, C = Critical Risk

III. Mitigation Strategies

For each identified and prioritized risk, develop specific mitigation strategies. These strategies should aim to reduce either the likelihood of the risk occurring or its potential impact, or both.

General mitigation strategies:

1. Privacy Breaches:
    - Implement strong encryption for data at rest and in transit
    - Use privacy-preserving AI techniques (e.g., federated learning, differential privacy)
    - Implement strict access controls and monitoring for AI systems
2. Bias and Discrimination:
    - Regularly test AI models for bias using diverse datasets
    - Implement fairness constraints in AI algorithms
    - Ensure diverse representation in AI development teams
3. Lack of Transparency:
    - Invest in explainable AI techniques
    - Develop clear communication protocols for AI use
    - Implement comprehensive logging and auditing systems
4. Overreliance on AI:
    - Maintain and update human expertise alongside AI development

- Implement human-in-the-loop processes for critical decisions
- Conduct regular manual audits of AI performance

5. AI Manipulation:
    - Implement adversarial training in AI model development
    - Regularly update and patch AI systems
    - Develop detection mechanisms for adversarial attacks
6. Ethical Decision-Making:
    - Establish clear ethical guidelines for AI systems
    - Implement ethics-aware AI algorithms
    - Ensure human oversight for ethically sensitive decisions
7. Job Displacement:
    - Develop reskilling and upskilling programs for cybersecurity staff
    - Create new roles focused on AI-human collaboration
    - Communicate transparently about AI implementation plans
8. Legal and Regulatory Compliance:
    - Regularly review and update AI systems to ensure compliance
    - Engage legal experts in AI development processes
    - Implement robust documentation and reporting procedures

For each specific risk, develop a detailed mitigation plan that includes:

- Specific actions to be taken
- Responsible parties
- Timeline for implementation
- Resources required
- Metrics for measuring effectiveness

Regular review and update of these mitigation strategies is crucial to ensure their ongoing effectiveness in the face of evolving AI technologies and cybersecurity threats.

2.2.5 Data Management

Effective and ethical data management is crucial for the successful implementation of GenAI in cybersecurity. This section outlines best practices for ensuring data quality, protection, and fairness.

I. Data Quality and Representativeness

To create accurate and unbiased GenAI models in cybersecurity, it is important to utilize quality representative data. Developing GenAI systems in cybersecurity require high-quality representative data to be the foundation of these solutions, underpinning unbiased and truly effective workflows. Organizations need to follow a full range of best practices to enable this. The process incorporates robust data quality assurance measures, such as audits and the definition of quality metrics and thresholds. It is vital to cover a broad spectrum of the cybersecurity landscape, including various network types and user's groups as well as threat scenarios. Since balanced datasets are crucial, especially for classification tasks (you can handle class imbalanced with the help of oversampling or under sampling techniques. Time-based features which capture the temporal aspects of cybersecurity data ensures that freshness is gained by keeping updated and well-preserved throughout regular update cycles. In practice, where no real-world data is available or the data must be kept private due to legal constraints, GenAI-techniques for generating XEC code/X-data will have been developed and well-established with different industries etc., but of course thoroughly validated next. Well defined data annotation, labelling and multiple annotators helps in reducing bias per person as well as maintain quality of the final dataset. Additionally, version control of datasets and documenting the changes helps in traceability also allows a better incremental improvement on AI models. If organizations follow these best practices, they can lay a strong foundation for their GenAI systems in cybersecurity to increase accuracy and fairness while also promoting reliability among AI-driven security solutions.

II. Data Protection Measures

In cybersecurity realm, GenAI systems require protection of sensitive data. The most basic element of data protection is strong encryption on the data and while in transit, all supported by secure key management practices. Again, this can be resolved by having strict role-based access control (RBAC) and multi-factor authentication along with periodic reviews of who has what level of permission. Protects Identification: Strong data anonymization techniques like k-anonymity, l-diversity and t-closeness keep users information personal. Well-secured, confined data storage areas and well-tested backup-and-recovery procedures protect against both attacks that steal or corrupt records. Attempting to limit data through the principles of data minimization, only taking and keeping in storage precisely what is needed by AI functions. Data at rest is safeguarded by secure data transfer protocols and these solutions also provide data loss prevention. Security — Policies regarding data destruction, especially with devices out of use that outline certified disposal methods prohibit unauthorized recovery. Federated

learning, homomorphic encryption and dierential privacy are some of the more advanced methods that have been developed to support AI model training while keeping individual data safe. These end-to-end safeguards are a part of critical data security to ensure that sensitive data used in GenAI cybersecurity systems is protected throughout its lifecycle, preserving the integrity and confidentiality of important information.

III. Addressing Bias in Training Data

By adding additional fairness constraints during the training of AI models, such as adversarial debiasing and fair representation learning bring up even more possibilities to diminish biased decisions from these models. Live monitoring outputs from AI models for bias, with feedback loops established Collect your own to check data quality and improve biases as you go through. In this context having diverse development teams even plays an important part in promoting a culture of inclusivity and questioning for possible biases. This lack of rigour in onboarding sources, particularly the approach to collect and record data transparently – doubts trustworthiness — transparency with the method for collecting it encourages accountability You can even add in some ethical data augmentation techniques to help over-represented classes and make sure we are augmenting all our targets alongside representation by group. The most exhaustive approach is to also periodically re-evaluate and incrementally update the training datasets, followed by assessments of effectiveness in AI model performance as well as fairness. Organizations following these data management approaches are better positioned to deploy GenAI for cybersecurity based on solid, transparent and ethical foundations that contribute towards the realization of fair AI-based security solutions.

2.2.6 Model Development and Deployment

Ethical considerations must be integrated throughout the entire lifecycle of GenAI model development and deployment in cybersecurity.

I. Explainable AI Techniques

Explainability — GenAI models must be explainable to build trust and ensure transparency in cybersecurity applications. This is done in a variety of ways and with thoughtful considerations. The first option is to use purely interpretable model architects — e.g. decision trees or rule-based systems, wherever possible; as well considering hybrid approaches that mix explainer-

based models with more complex architecture and produce the final metrics using only explainable features/actions. For black-box model, post-hoc explainability methods such as LIME or SHAP can be followed with importance analysis to find the features that cause a decision. Deep learning is in vogue and adding an attention mechanism to these models can make certain key input features stick out improving its interpretability. Counterfactual explanations shed light on what changes would trigger the model decisions to differ. Specific model techniques like attention visualization for language models or anomaly explanations in detection models give explainability, intended to the usage at hand. It is also looking for explanation interfaces to be user-friendly, while being tailored depending on the stakeholders involved. Finally, setting up quality metrics for explanations and collecting user feedback guarantees that explainability is forever enriched. Together, this makes GenAI models more interpretable in cybersecurity and builds trust to facilitate increased human—Artificial Intelligence (AI) collaboration across security operations.

II. Testing and Validation Procedures

III. Monitoring and Auditing Processes

Ongoing monitoring and auditing are essential for maintaining the ethical performance of deployed GenAI systems in cybersecurity. This process involves real-time monitoring of AI system performance and outputs, with alert mechanisms for anomalous behavior. Key performance indicators related to accuracy, fairness, and efficiency are tracked continuously. Bias detection mechanisms are implemented to identify emerging biases or discrimination in AI outputs. Comprehensive, tamper-proof audit logs of AI decisions and actions are maintained for investigative purposes. Regular ethical audits assess compliance with guidelines and identify areas for improvement. Drift detection monitors for concept or data drift that may affect model performance, triggering automatic retraining or alerts when necessary. User feedback channels are established to identify potential issues or improvements. External audits and participation in industry-wide benchmarking programs provide independent verification of ethical compliance. The effectiveness of AI systems is monitored during real cybersecurity incidents, with post-incident reviews conducted to assess performance and identify lessons learned. Regular transparency reports on AI system performance and ethical compliance are produced and shared with stakeholders. These practices collectively ensure that GenAI systems

in cybersecurity remain effective, ethically sound, and trustworthy throughout their deployment.

2.2.7 Transparency and Explainability

Ensuring transparency and explainability in GenAI systems used for cybersecurity is crucial for building trust, enabling effective human oversight, and meeting ethical and legal requirements.

I. Documenting AI Decision-Making

Comprehensive documentation of AI decision-making processes is essential for transparency and accountability in GenAI systems for cybersecurity. This includes detailed documentation of model architectures, training data sources and preparation processes, and decision logic underlying AI operations. Version control for AI models and associated documentation ensures a clear history of changes and updates. Key performance metrics, including accuracy, precision, recall, and fairness measures, are documented along with their calculation methods. Known limitations, boundary conditions, and scenarios requiring human intervention are clearly specified. Ethical considerations, including fairness measures and bias mitigation strategies, are documented. Regulatory compliance documentation outlines how the AI system meets relevant requirements and includes any certifications or assessments. Incident response documentation provides guidance on handling AI system failures and investigating AI-related issues. This comprehensive approach to documentation enhances transparency, facilitates accountability, and supports effective management of GenAI systems in cybersecurity throughout their lifecycle.

II. Stakeholder Communication

Effective communication with stakeholders about AI systems is crucial for maintaining transparency and trust.

Strategies for stakeholder communication:

1. Layered Explanations:
    - Develop explanations at different levels of technical detail for various stakeholders

- Provide high-level overviews for non-technical stakeholders and detailed technical documentation for specialists

2. Regular Reporting:
    - Establish a schedule for regular reporting on AI system performance and impact
    - Include both quantitative metrics and qualitative assessments
3. Interactive Dashboards:
    - Develop user-friendly dashboards for real-time monitoring of AI system performance
    - Allow stakeholders to explore key metrics and explanations
4. Explainable AI Interfaces:
    - Implement interfaces that allow users to query AI decisions and receive explanations
    - Ensure explanations are contextual and relevant to the user's role and expertise
5. Training and Education:
    - Provide training sessions for stakeholders on how to interpret AI outputs and explanations
    - Develop educational materials on the basics of AI in cybersecurity
6. Feedback Channels:
    - Establish clear channels for stakeholders to provide feedback or raise concerns about AI systems
    - Ensure timely responses to stakeholder inquiries and concerns
7. Transparency in AI Use:
    - Clearly communicate when and where AI is being used in cybersecurity processes
    - Provide information on the role of AI in decision-making and its limitations
8. Case Studies and Examples:
    - Develop case studies illustrating how AI systems work in real-world cybersecurity scenarios
    - Use concrete examples to explain complex AI concepts and decision processes

III. Audit Trail Maintenance

Maintaining a comprehensive audit trail is crucial for accountability and transparency in AI-driven cybersecurity systems. This involves logging all significant AI decisions, including timestamps, input data, outputs, and confidence levels. Model invocation records track each instance an AI model is used, noting the version and context. Data access logs document all data interactions, while user interaction logs record human-AI exchanges and any manual overrides. System changes and updates are meticulously recorded, including model retraining and parameter adjustments. Performance monitoring logs track ongoing metrics and anomalies. Incident response logs detail AI system reactions to security events and any human interventions. Compliance check logs document adherence to ethical guidelines and regulations. All logs are securely stored with appropriate access controls and retention policies. Tools for easy searching and analysis of audit logs are implemented, enabling cross-system correlation. Regular reviews of audit trails help identify patterns and inform AI system improvements. This comprehensive approach ensures that the use of GenAI in cybersecurity remains accountable, trustworthy, and ethically aligned.

2.2.8 Human-AI Collaboration

Effective collaboration between humans and AI systems is crucial for ethical and efficient cybersecurity operations. This section outlines best practices for fostering productive human-AI partnerships.

I. Human Oversight Levels

Establishing appropriate levels of human oversight is essential for maintaining control and accountability in AI-driven cybersecurity systems.

Levels of human oversight:

1. AI-Assisted Decision Making:
    - AI provides recommendations, but humans make final decisions
    - Suitable for high-stakes or ethically sensitive situations
2. Human-in-the-Loop:
    - AI makes decisions, but requires human approval before actions are taken
    - Appropriate for medium-risk scenarios or when building trust in AI systems

3. Human-on-the-Loop:
    - AI operates autonomously, but humans monitor and can intervene if necessary
    - Suitable for low-risk, high-volume tasks where quick response is crucial
4. Fully Autonomous:
    - AI operates without direct human oversight
    - Appropriate only for well-defined, low-risk tasks with clear boundaries

Implementation strategies:

- Conduct risk assessments to determine appropriate oversight levels for different tasks
- Implement technical safeguards to enforce required human involvement
- Regularly review and adjust oversight levels based on AI system performance and evolving risks

II. AI-Assisted Decision-Making Protocols

Developing clear protocols for AI-assisted decision-making is essential for consistent and ethical use of AI in cybersecurity operations. These protocols should clearly define the roles of AI systems and human operators, specifying which decisions can be made autonomously and which require human input. Confidence thresholds for AI recommendations and criteria for mandatory human review should be established. Clear escalation procedures based on risk and complexity are necessary, along with specific protocols for time-critical situations and fallback procedures when immediate human oversight is unavailable. Processes for resolving disagreements between AI recommendations and human judgments should be developed, with a framework for weighing AI insights against human expertise. Comprehensive documentation requirements ensure transparency in decision-making processes. Finally, feedback loops should be implemented to allow human operators to provide input on AI recommendations, facilitating continuous improvement of AI system performance and decision-making protocols. These elements collectively ensure a balanced, efficient, and ethically sound approach to AI-assisted decision-making in cybersecurity operations.

III. Staff Training for AI Collaboration

Comprehensive training is crucial for enabling effective human-AI collaboration in cybersecurity. This training should cover AI literacy, providing foundational knowledge on AI concepts, capabilities, strengths, and limitations in cybersecurity contexts. Staff must be educated on ethical considerations, including potential biases and fairness issues in AI systems. Training should focus on interpreting AI outputs, teaching critical evaluation of AI recommendations and explainable AI techniques. Human-AI interaction skills, including when to override or question AI decisions, are essential. Detailed training on specific AI systems used in the organization, including interfaces and operational procedures, is necessary. Scenario-based training using realistic cybersecurity tasks helps practice human-AI collaboration in both routine and crisis situations. Continuous learning programs keep staff updated on AI advancements, while cross-functional understanding fosters collaboration between AI specialists and domain experts. Ethical decision-making frameworks and case studies of ethical dilemmas in AI-driven cybersecurity should be discussed.

Finally, staff should be trained to provide constructive feedback on AI system performance, fostering a culture of continuous improvement in human-AI collaboration. These practices create a synergistic relationship between human expertise and AI capabilities, enhancing the effectiveness and ethical standing of cybersecurity operations.

2.2.9 Continuous Improvement

Maintaining an ethical and effective GenAI system in cybersecurity requires ongoing efforts to assess, update, and improve practices. This section outlines key strategies for continuous improvement.

I. Ethical Audits and Assessments

Regular ethical audits and assessments are crucial for ensuring ongoing compliance with ethical standards and identifying areas for improvement in AI-driven cybersecurity systems. These audits should be scheduled regularly, incorporating both internal and external independent assessments. The audit scope should cover compliance with ethical guidelines, evaluating fairness, transparency, accountability, and privacy aspects, as well as reviewing data management practices and model performance. Stakeholder feedback from end-users, cybersecurity teams, and management provides valuable insights into the perceived ethical standing of AI systems. Technical assessments, including evaluations for bias, explainability, and robustness, along with penetration testing, help identify potential vulnerabilities.

Documentation reviews ensure completeness and accuracy of AI system records. Compliance checks verify adherence to relevant regulations, industry standards, and internal ethical policies. Incident reviews analyze ethical issues or near-misses, assessing the effectiveness of response and mitigation measures. Benchmarking against industry best practices helps identify areas of leadership or improvement in ethical AI implementation. This comprehensive approach to ethical auditing ensures continuous improvement and maintains the integrity of AI systems in cybersecurity.

II. Keeping Up with AI Ethics Guidelines

The field of AI ethics is rapidly evolving, and organizations must stay informed about new guidelines, standards, and best practices.

Strategies for staying current:

1. Dedicated Ethics Team:
    - Assign responsibility for monitoring AI ethics developments to a dedicated team or individual
    - Ensure regular reporting on relevant updates to key stakeholders
2. Industry Partnerships:
    - Participate in industry groups and forums focused on AI ethics in cybersecurity
    - Engage in collaborative efforts to develop and refine ethical standards
3. Academic Collaboration:
    - Establish partnerships with academic institutions researching AI ethics
    - Participate in or sponsor relevant research projects
4. Regulatory Monitoring:
    - Implement processes to monitor and interpret new regulations related to AI ethics
    - Engage with regulatory bodies to understand and influence upcoming guidelines
5. Professional Development:
    - Encourage staff to attend conferences, workshops, and training sessions on AI ethics
    - Support certifications in AI ethics for relevant team members
6. Subscription Services:

- Subscribe to reputable AI ethics newsletters and update services
- Utilize AI-powered monitoring tools to track relevant publications and discussions

7. Internal Knowledge Sharing:
    - Establish channels for sharing and discussing new AI ethics developments within the organization
    - Conduct regular briefings or seminars on emerging ethical considerations
8. Ethical AI Framework Review:
    - Regularly review and update the organization's ethical AI framework
    - Incorporate new guidelines and best practices as they emerge

III. Feedback Integration

Establishing robust feedback mechanisms and effectively integrating feedback is crucial for continuous improvement of ethical AI practices in cybersecurity. This process involves implementing multi-channel feedback collection methods and encouraging ongoing input from all stakeholders. A structured analysis approach, using tools like text analytics and sentiment analysis, helps process and categorize large volumes of feedback. A prioritization framework ensures that high-priority ethical concerns are addressed promptly. Cross-functional teams should be involved in reviewing and acting on feedback, promoting collaborative problem-solving for complex ethical issues. Continuous monitoring of AI system performance based on user feedback, coupled with alert mechanisms for recurring or severe ethical issues, enables rapid response. Feedback-driven updates should be incorporated into AI model retraining, with ethical considerations central to system modifications. Transparency in responding to feedback, including clear communication of actions taken, builds trust with stakeholders. Closing the feedback loop by following up with providers and seeking input on implemented changes ensures effectiveness. Documenting lessons learned informs future AI development and deployment. Incentivizing ethical feedback fosters a culture of ethical vigilance and continuous improvement. These practices collectively ensure that GenAI use in cybersecurity remains ethically sound, effective, and aligned with evolving best practices and stakeholder expectations.

2.2.10 Incident Response and Accountability

Effective incident response and clear accountability mechanisms are crucial for maintaining trust and ethical standards in AI-driven cybersecurity systems.

I. AI-Related Ethical Incident Protocols

Developing specific protocols for handling AI-related ethical incidents is crucial for ensuring rapid and appropriate responses to ethical breaches or concerns in cybersecurity. These protocols should include a comprehensive incident classification system with established severity levels, and clear reporting mechanisms that offer anonymity to encourage reporting without fear of retaliation. Procedures for initial assessment should be defined, including criteria for escalation to higher management or external authorities. Containment strategies should be developed to quickly limit the impact of ethical incidents, including steps for temporarily suspending AI systems if necessary. A thorough investigation process should be outlined, emphasizing evidence preservation and integrity. Mitigation and remediation guidelines should address both short-term fixes and long-term solutions for the root causes of ethical incidents. Protocols for timely and transparent stakeholder communication, tailored to incident severity and impact, are essential. Comprehensive documentation and reporting procedures, including templates for incident reports and post-incident analyses, should be implemented. Continuous monitoring processes should be established to detect similar or recurring issues, with feedback loops to improve incident detection and response. Regular training sessions and simulations should be conducted to practice handling various ethical scenarios, ensuring team readiness for real-world incidents. This comprehensive approach to AI ethical incident management helps maintain the integrity and trustworthiness of AI systems in cybersecurity operations.

II. Responsibility and Accountability

Clearly defined responsibility and accountability structures are essential for ethical AI governance in cybersecurity. This framework begins with clear role definitions and a well-established chain of command for ethical decision-making and incident response. Overall responsibility for AI ethics should be assigned to a senior executive, such as a Chief Ethics Officer, with the authority and resources to enforce ethical standards. Ethical responsibilities should be defined at both team and individual levels, incorporating ethical considerations into performance evaluations and goals. A comprehensive ethical decision-making framework should guide AI-related matters, providing clear escalation procedures for ethical concerns. Strong whistleblower protection policies are crucial to encourage reporting of ethical issues

without fear of retaliation. Continuous education ensures all staff understand their roles in maintaining ethical AI systems. External accountability involves engaging with stakeholders and being transparent about accountability measures. Regular audits and reviews of ethical compliance and accountability structures maintain their effectiveness. Finally, clear consequences for ethical breaches, applied fairly and consistently, reinforce the importance of ethical conduct. This robust accountability structure ensures that ethical considerations are deeply embedded in all aspects of AI use in cybersecurity, promoting responsible and trustworthy AI implementation.

III. Stakeholder Communication During Incidents

Effective communication with stakeholders during AI-related ethical incidents is crucial for maintaining trust and transparency in cybersecurity operations. This process begins with timely notification protocols, establishing clear timeframes for initial communication based on incident severity. Transparent disclosure is essential, providing honest information about the incident's nature without downplaying its severity or potential impact. Regular updates should be scheduled throughout the incident response process, keeping stakeholders informed of progress, challenges, and next steps. Communication strategies should be tailored to different stakeholder groups, ensuring appropriate technical levels for each audience. A multi-channel approach helps reach all stakeholders consistently. Q&A and feedback mechanisms allow stakeholders to seek clarification and provide input, with prompt and thorough responses. Post-incident communication should include comprehensive reports, lessons learned, and preventive measures. Senior leadership involvement in serious incidents demonstrates organizational commitment to ethical concerns. All communications must comply with legal and regulatory requirements, with legal counsel consulted for serious incidents. Lastly, communications should convey empathy, acknowledging the impact on affected parties, while taking responsibility and focusing on constructive solutions. These strategies collectively enable organizations to effectively manage AI-related ethical incidents, maintain stakeholder trust, and demonstrate commitment to ethical AI practices in cybersecurity.

2.2.11 Stakeholder Engagement

Effective stakeholder engagement is crucial for the successful and ethical implementation of GenAI in cybersecurity. This section outlines strategies for engaging with various stakeholders throughout the AI lifecycle.

I. Engaging Relevant Stakeholders

Identifying and engaging with all relevant stakeholders is crucial for ensuring comprehensive input and buy-in for AI initiatives in cybersecurity. This process begins with thorough stakeholder mapping, identifying key groups such as employees, customers, partners, and regulators, and assessing their interests and potential concerns. A structured engagement plan should be developed, defining objectives, methods, and frequency of interaction for each stakeholder group. Creating forums for inclusive dialogue on AI ethics in cybersecurity encourages diverse perspectives and proactive problem-solving. Involving key stakeholders in decision-making processes for AI initiatives, particularly in developing ethical guidelines and implementation strategies, fosters a sense of ownership and commitment. Regular updates on AI developments and ethical considerations should be provided through various channels. Robust feedback mechanisms ensure stakeholder input is systematically collected, analyzed, and acted upon. Transparency initiatives, sharing information about AI use cases, benefits, and potential risks, build trust and understanding. Stakeholder education, including training sessions on AI ethics, helps create a well-informed ecosystem. Building partnerships with academic institutions, industry bodies, and ethics organizations facilitates collaboration on research and best practices. Finally, engaging with the broader community on AI ethics in cybersecurity contributes to the societal dialogue on AI. These strategies collectively ensure a comprehensive, inclusive approach to stakeholder engagement in AI initiatives, promoting ethical and effective implementation in cybersecurity.

II. AI Ethics Education and Training

Comprehensive education and training on AI ethics is essential for all stakeholders involved in or affected by AI-driven cybersecurity systems. This training should start with foundational AI knowledge, covering basic concepts, capabilities, and limitations, and explaining AI's application in cybersecurity. Core ethical principles relevant to AI in cybersecurity should be taught, using real-world examples and case studies to illustrate ethical dilemmas. Education on bias and fairness is crucial, including methods for detecting and mitigating bias in AI systems. Privacy and data protection training should cover relevant regulations and best practices. The importance of transparency and explainability in AI systems should be emphasized, along with techniques for interpreting AI decisions. Training on effective human-AI collaboration and the role of human judgment in AI-assisted decision-making is vital. Ethical decision-making frameworks should be provided and practiced through role-playing exercises. Incident response training, including simulations of ethical breaches, prepares stakeholders for real-

world scenarios. The importance of ongoing learning in AI ethics should be stressed, with resources provided for staying updated. Finally, role-specific training modules should address the unique ethical considerations relevant to different positions within the organization. This comprehensive approach to AI ethics education ensures that all stakeholders are well-equipped to navigate the ethical challenges of AI in cybersecurity.

III. Feedback Channels

Establishing effective feedback channels is crucial for ongoing communication and improvement in AI ethics practices within cybersecurity. A multi-channel approach, including surveys, suggestion boxes, and ethics hotlines, ensures accessibility for all stakeholder groups. Anonymous reporting options encourage open communication and protect whistleblowers reporting ethical concerns. Regular surveys on AI ethics perceptions help inform policy updates and training programs. Stakeholder forums or town halls facilitate open dialogue on AI ethics issues. A clear feedback integration process demonstrates how input influences AI ethics policies and practices. Responsive communication, with timely updates on actions taken, maintains stakeholder engagement. Incentivizing feedback creates a culture valuing ethical vigilance and open communication. User experience monitoring helps identify potential ethical issues or areas for improvement. External audits assess the effectiveness of stakeholder engagement, providing insights for process improvement. Continuous improvement of feedback channels, based on stakeholder input, ensures they remain relevant and effective. This comprehensive approach to feedback channels ensures that an organization's AI ethics practices in cybersecurity remain inclusive, responsive, and aligned with stakeholder expectations, enhancing both the ethical standing and effectiveness of AI initiatives.

2.3 Risk Assessment and Mitigation Plan

2.3.1 Cyber Security Risk Assessment

Risk Assessment of GenAI application in cybersecurity for Automobile Service Industry is conducted following NIST Special Publication 800-30 Revision 1, titled "Guide for Conducting Risk Assessments," which offers a structured framework for evaluating risks to information systems (Blank and Gallagher 2012). The guide emphasizes a systematic approach

starting with the preparation phase, where organizations define the methodology of the risk assessment. This involves identifying the information system boundaries and assets at risk (Fikri et al. 2019). During the assessment phase, identify and analyse potential threats and vulnerabilities, assessing their likelihood and impact to determine overall risk levels. The process includes identifying threats such as cyber-attacks or insider threats and vulnerabilities like unpatched software or misconfigured systems. (Safitri and Kabetta 2023). The guide suggests categorizing likelihood and impact to evaluate the risk comprehensively.

The next steps involve selecting and implementing appropriate controls to mitigate identified risks. NIST CSF 2.0 core functions provides the framework which is used as cybersecurity controls to mitigate Risks. The CSF Core Functions GOVERN, IDENTIFY, PROTECT, DETECT, RESPOND, and RECOVER organize cybersecurity outcomes at their highest level. The NIST Cybersecurity Framework (CSF) outlines six core functions designed to help organizations manage and respond to cybersecurity risks comprehensively. The first function, **Govern (GV)**, focuses on establishing and communicating the organization's cybersecurity risk management strategy, policies, and governance structures. This function ensures that roles and responsibilities are clearly defined and that cybersecurity efforts align with the organization's broader enterprise risk management (ERM) strategy.

Next, the **Identify (ID)** function helps organizations gain a clear understanding of their current cybersecurity risks by identifying critical assets such as data, hardware, software, systems, and people. By understanding these assets and their associated risks, organizations can prioritize their cybersecurity efforts to address the most pressing vulnerabilities.

The **Protect (PR)** function involves implementing safeguards to manage the identified risks, helping to prevent or reduce the likelihood of cybersecurity incidents. These safeguards include measures like access control, identity management, and data security, along with training programs to raise awareness among employees.

The **Detect (DE)** function ensures that organizations have the capabilities to identify and analyze potential cybersecurity threats or incidents in a timely manner. By monitoring for anomalies and indicators of compromise, organizations can quickly detect attacks before they cause significant damage.

The **Respond (RS)** function comes into play, guiding organizations in taking action to contain and mitigate the impact of cybersecurity incidents. This function involves conducting detailed

incident analysis, managing the containment process, and ensuring that the incident is communicated appropriately to stakeholders. The goal is to minimize the damage caused by the incident and to take steps to prevent its recurrence.

Finally, the **Recover (RC)** function focuses on restoring affected assets and operations following a cybersecurity incident. Recovery efforts aim to quickly return to normal business operations while reducing the long-term impact of the event. This function also emphasizes clear communication during the recovery process to ensure that all stakeholders are informed, and that the organization can continue to operate smoothly in the aftermath of an incident.

These controls can be technical, administrative, or physical and should be continuously monitored to ensure their effectiveness. Regular reviews and adjustments are crucial as the risk environment evolves (Möller 2023). Overall, NIST SP 800-30 Rev. 1 and NIST CSF 2.0 provides a detailed methodology for identifying, analysing, and managing risks, helping organizations protect their information systems and data effectively. In cybersecurity, understanding the relationships between threats, vulnerabilities, and assets is essential for developing robust protection strategies. The Threat-Vulnerability-Asset (TVA) model helps in identifying and managing these elements systematically.

Threats are potential events or actions that can exploit vulnerabilities to cause harm or damage to an organization. They represent the external or internal forces that pose risks. Examples include cyberattacks, insider threats, and natural disasters. (Asaad and Saeed 2022)

Vulnerabilities are weaknesses or flaws within a system that can be exploited by threats. They represent gaps in security that could be used to breach or damage the system. Vulnerabilities can be due to software bugs, misconfigurations, lack of updates, or weak authentication mechanisms. (Asaad and Saeed 2022)

Assets are the resources or components of an organization that need protection. They can include hardware, software, data, and processes. Assets are valuable because they store, process, or transmit sensitive information or are crucial for operations. (Asaad and Saeed 2022)

2.3.2 TVA in Cybersecurity:

Threat-Vulnerability-Asset Relationships:

1. Threat: A cyberattack aiming to exploit a weakness.
2. Vulnerability: The weakness in the system, such as unpatched software.

3. Asset: The system or data that the threat is targeting.

For example, if a threat is a malware attack, the vulnerability could be an unpatched software, and the asset could be a server storing critical data.

Using TVA for Risk Management:

1. Identify Threats: Understand what potential threats exist (e.g., phishing attacks, data breaches).
2. Assess Vulnerabilities: Identify weaknesses that could be exploited by these threats (e.g., weak passwords, outdated software).
3. Protect Assets: Focus on securing assets that are at risk (e.g., sensitive data, critical systems).

TVA Example:

1. Threat: Software attack (e.g., code injection).
2. Vulnerability: Poorly secured API endpoints.
3. Asset: GenAI application that processes and stores customer data.

In this case, the threat of a software attack could exploit the vulnerability of poorly secured API endpoints to compromise the asset, which is the GenAI application. This would lead to potential data breaches and operational disruption.

Understanding these relationships, risk analysis is conduct and organizations can prioritize security efforts to address the most critical vulnerabilities and protect their valuable assets from potential threats. This systematic approach helps in developing effective risk management and mitigation strategies (Blank and Gallagher 2012). Table below shows the Assets which are associated with the organization.

2.3.3 Assets Table

Asset No.	Asset Type	Asset
1.	**Hardware**	Servers (Cloud-based) that host the GenAI application and website
		Web Server for hosting the company's website
		Routers for network connectivity
		Firewalls to protect the network

		Switches to connect network devices
		Workstations/Computers used by staff
		Telecommunications Equipment (phones, intercoms)
		Load Balancers (Cloud-based) to manage website traffic
		Point of Sale (POS) Systems for handling customer payments
		Carwash Equipment (automated carwash machines, sensors, etc.)
2.	**Software**	GenAI Application
		Content Management System (CMS) for website updates
		Operating Systems on servers and workstations
		Security Tools (e.g., antivirus, threat detection software)
		APIs for integrating GenAI with the website and other systems
		SSL/TLS Certificates for secure web communication
		Point of Sale (POS) Software
		Customer Relationship Management (CRM) Software
		Business Management Software (inventory, scheduling, etc.)
		Monitoring Tools for network and application performance
		Logging Systems for tracking activities and troubleshooting
3.	**Data**	Training Data for the GenAI model
		Customer Data (personal details, service history, payment info)
		Operational Data (logs, configurations, business metrics)
		Proprietary Information (business strategies, pricing models)
		Financial Data (transaction records, financial statements)
		Compliance Documentation Data (data protection, safety regulations)
		Legal Agreements Data (customer terms, supplier contracts, privacy policies)
4.	**Process**	Network Setup Process
		Device Configuration Process
		GenAI Tool Development and Deployment Process
		Customer Interaction Process (in-person and online)
		Service Booking Process (online and on-site)

		Payment Processing Process
		Change Management Process
		Incident Response Process
		Data Management Process
		Access Control Process for IT systems and physical premises
		Employee Training Process (for using IT systems and carwash equipment)
		Disaster Recovery Plan for IT infrastructure and carwash operations process
5.	**Cloud Services**	Cloud Storage for data backups
		Cloud Compute Instances for running the GenAI model
		Cloud Security Services to protect cloud-based resources

Table-5 Assets Table

2.3.4 Threat Table

Threat No	Threat Name	Description	Related Assets
1	**Software Attacks**	Includes Malware, DoS (Denial of Service), DDoS (Distributed Denial of Service), MITM (Man in the Middle), Sniffing, DNS Poisoning, Spoofing, Code Injection, Backdoors, and Ransomware.	GenAI Application, Operating System, Security Tools, APIs, Servers (Cloud-based)

2	**Human Errors**	Refers to Phishing, Server Misconfiguration, Social Engineering, and Accidental Data Deletion.	All Assets (Hardware, Software, Data, Processes)
3	**Trespass**	Unauthorized Access via password attacks, such as Dictionary Attack, Brute Force Attack, Shoulder Surfing, and Physical Tampering.	Servers, Router, Switch, Computer, Customer Data, Access Control Process, APIs
4	**Information Extortion**	Ransomware attacks leading to Blackmail or Information Disclosure for Financial Gain.	Customer Data, Proprietary Information, Operational Data, GenAI Application
5	**Hardware Failures/Errors**	Includes Drive Failures, Server Crashes, and Network Hardware Malfunctions.	Servers, Router, Firewall, Switch, Computer
6	**Wi-Fi Eavesdropping**	Capturing Sensitive Information, Passwords, or Other Confidential Data through Unauthorized Wireless Network Monitoring.	Router, Switch, Access Control Process, Customer Data
7	**Software Failures/Errors**	Bugs, Code Performance Issues, Loopholes, and Unpatched Vulnerabilities.	GenAI Application, Operating System, APIs, Security Tools, Chatbot Development Process, Service Booking Process
8	**Sabotage and Vandalism**	Deliberate Destruction or Tampering of Assets, including Physical Damage to Hardware and Digital Destruction (e.g., Data Wiping, Defacement).	Servers, Firewall, Proprietary Information, Data Management Process
9	**Forces of Nature**	Natural Disasters such as Cyclones, Fires, Floods, and Lightning Strikes, leading to Data Loss, Hardware Damage, or Operational Downtime.	All Physical Hardware (Servers, Router, Firewall, Switch, Computer), Training Data, Customer Data

10	**Data Breaches**	Unauthorized access to sensitive data, resulting in information leakage and compliance violations.	Customer Data, Proprietary Information, Training Data, Operational Data
11	**Insider Threats**	Malicious actions taken by employees or contractors, such as data theft, sabotage, or misuse of access privileges.	All Assets (Hardware, Software, Data, Processes)
12	**Supply Chain Attacks**	Attacks targeting the organization's suppliers, leading to compromised components or software used within the infrastructure.	GenAI Application, APIs, Security Tools, Operating System
13	**API Abuse**	Unauthorized use or exploitation of API endpoints, leading to data breaches, service disruption, or unauthorized actions.	APIs, GenAI Application, Operating System
14	**Credential Theft**	Theft of authentication credentials through phishing, malware, or brute-force attacks, leading to unauthorized access.	Servers, Router, Firewall, Switch, GenAI Application, Customer Data, Access Control Process
15	**Cloud Security Misconfigurations**	Improper configuration of cloud services, leading to exposure of sensitive data or unauthorized access to cloud-based systems.	Servers (Cloud-based), GenAI Application, Customer Data, Proprietary Information
16	**Zero-Day Exploits**	Exploitation of unknown vulnerabilities in software before they can be patched, leading to unauthorized access or system compromise.	Operating System, GenAI Application, APIs, Security Tools
17	**Regulatory Non-Compliance**	Failure to comply with relevant data protection and cybersecurity	All Data (Customer Data, Training Data, Operational Data,

		regulations, leading to fines, legal action, and reputational damage.	Proprietary Information), Processes related to Data Management and Compliance
18	**Data Poisoning**	Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs.	GenAI Application, Training Data
19	**Model Inversion**	Attacker's reverse-engineering the model to infer sensitive information from the training data.	GenAI Application, Training Data
20	**Adversarial Attacks**	Small, crafted inputs causing the GenAI model to make incorrect or harmful predictions.	GenAI Application
21	**Model Drift**	Performance degradation over time as the model encounters data differing from its training set.	GenAI Application
22	**Hallucination**	Generation of nonsensical or incorrect outputs that appear plausible.	GenAI Application
23	**Bias and Discrimination**	Inherited biases from training data leading to unfair or discriminatory outputs.	GenAI Application
24	**Explainability**	Difficulty in understanding and interpreting model outputs, obscuring errors and reducing trust.	GenAI Application
25	**Dependency on Training Data Quality**	Inadequate or biased training data leading to flawed models and erroneous outputs.	GenAI Application
26	**Intellectual Property Violations**	Unintentional plagiarism or copyright infringement due to the replication of protected content.	GenAI Application, Proprietary Information

27	**Overfitting**	The model performs well on training data but poorly on new, unseen data.	GenAI Application
28	**Malicious Use of GenAI**	GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks.	GenAI Application
29	**Privacy Invasion**	The application inadvertently exposing sensitive personal or organizational data through its outputs.	GenAI Application, Customer Data
30	**Algorithmic Manipulation**	Attackers manipulating the GenAI's algorithm to produce biased or harmful outputs.	GenAI Application
31	**Unauthorized Access**	Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs.	GenAI Application
32	**Ethical Violations**	GenAI producing outputs that violate ethical standards, such as generating offensive content.	GenAI Application

Table-6 Threats Table

2.3.5 Vulnerability

Vulnerabilities are weaknesses or flaws within a system that can be exploited by threats to cause harm or unauthorized access. In the context of Implementing GenAI application for the Automobile Service Industry several critical vulnerabilities have been identified and are in table below.

Vulnerability Table

Threat	Asset	Vul. No	TVA	Vulnerability
Software Attacks	Firewall	1	T1V1A1	Weak authentication, such as easily guessable passwords.
Software Attacks	GenAI Application	2	T1V2A2	Unpatched vulnerabilities in the GenAI codebase.
Software Attacks	APIs	3	T1V3A3	Poorly secured API endpoints prone to injection attacks.
Human Errors	Server (Cloud-based)	4	T2V4A1	Misconfiguration during server setup.
Human Errors	Operating System	5	T2V5A2	Lack of timely updates and patches.
Trespass	Router	6	T3V6A1	Default credentials not changed, leading to unauthorized access.
Information Extortion	Customer Data	7	T4V7A3	Lack of encryption, leading to data exposure during extortion.
Hardware Failures/Errors	Servers (Cloud-based)	8	T5V8A1	Single point of failure in server hardware.
Wi-Fi Eavesdropping	Switch	9	T6V9A1	Insufficient encryption on wireless communication.
Software Failures/Errors	GenAI Application	10	T7V10A2	Bugs and code performance issues leading to system crashes.
Sabotage and Vandalism	Servers (Cloud-based)	11	T8V11A1	Lack of physical security, leading to tampering.
Forces of Nature	All Physical Hardware	12	T9V12A1	Lack of disaster recovery plans for natural disasters.

Data Breaches	Customer Data	13	T10V13A3	Weak access controls leading to unauthorized data access.
Insider Threats	All Assets	14	T11V14A1-5	Insufficient monitoring and auditing of employee activities.
Supply Chain Attacks	GenAI Application	15	T12V15A2	Compromised third-party software or components used within the application.
API Abuse	APIs	16	T13V16A2	Insufficient rate limiting and authentication mechanisms on APIs.
Credential Theft	Access Control Process	17	T14V17A4	Weak password policies leading to credential theft.
Cloud Security Misconfigurations	Servers (Cloud-based)	18	T15V18A1	Improper configuration of cloud services, leading to exposure of sensitive data.
Zero-Day Exploits	Operating System	19	T16V19A2	Unpatched zero-day vulnerabilities exploited by attackers.
Regulatory Non-Compliance	Customer Data, Training Data, Proprietary Information	20	T17V20A3	Lack of adherence to data protection and cybersecurity regulations, leading to legal and financial penalties.
Data Poisoning	GenAI Application	21	T18V21A2	Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs.
Model Inversion	GenAI Application	22	T19V22A2	Attacker's reverse-engineering the model to

				infer sensitive information from the training data.
Adversarial Attacks	GenAI Application	23	T20V23A2	Small, crafted inputs causing the GenAI model to make incorrect or harmful predictions.
Model Drift	GenAI Application	24	T21V24A2	Performance degradation over time as the model encounters data differing from its training set.
Hallucination	GenAI Application	25	T22V25A2	Generation of nonsensical or incorrect outputs that appear plausible.
Bias and Discrimination	GenAI Application	26	T23V26A2	Inherited biases from training data leading to unfair or discriminatory outputs.
Dependency on Training Data Quality	GenAI Application	28	T24V27A2	Inadequate or biased training data leading to flawed models and erroneous outputs.
Intellectual Property Violations	Proprietary Information	29	T25V28A3	Unintentional plagiarism or copyright infringement due to the replication of protected content.
Overfitting	GenAI Application	30	T26V29A2	The model performs well on training data but poorly on new, unseen data.
Malicious Use of GenAI	GenAI Application	31	T27V30A2	GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks.

Privacy Invasion	GenAI Application and Data	32	T28V31A3	The application inadvertently exposing sensitive personal or organizational data through its outputs.
Algorithmic Manipulation	GenAI Application	33	T29V32A2	Attackers manipulating the GenAI's algorithm to produce biased or harmful outputs.
Unauthorized Access	All Assets	34	T30V33A1-5	Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs.
Ethical Violations	GenAI Application	35	T31V34A2	GenAI producing outputs that violate ethical standards, such as generating offensive content.

Table-7 Vulnerability Table

2.3.6 Risk Rank

In risk assessment, the risk rank is a crucial metric used to prioritize vulnerabilities based on their likelihood of exploitation and the potential impact if they are exploited. The likelihood measures how probable it is that a vulnerability will be exploited, categorized as High, Moderate, or Low. Impact evaluates the potential consequences of exploitation, also categorized as High, Moderate, or Low. By combining these two factors, the overall risk is determined (Blank and Gallagher 2012).

For example, a vulnerability with High likelihood and High impact is deemed a High-risk issue, requiring immediate attention and remediation. Conversely, a vulnerability with Low likelihood and Low impact is considered Low risk and is given a lower priority. The risk rank assigns a numerical value or category, such as Rank 1 for Critical risks that demand urgent action, rank 2 for Moderate risks that can be tackled after more pressing issues. Lower ranks,

like Rank 3 correspond to risks that are less critical and can be addressed as resources permit. This prioritization ensures that the most significant threats are managed first, effectively reducing the potential for substantial damage or disruption to the organization.

Likelihood	Impact	Combined Risk	Risk
High	Low	High	Low
Moderate	Low	Moderate	Low
Low	Low	Low	Low
High	Moderate	High	Moderate
Moderate	Moderate	Moderate	Moderate
Low	Moderate	Low	Low
High	High	High	High
Moderate	High	Moderate	Moderate
Low	High	Low	Low

Table-8 Risk Rank Value Table

2.3.6.1 Vulnerability occurrence and Risk Assessment Table

Vul No	Vulnerability	Likelihood	Impact	Risk	Rank
1	Weak authentication, such as easily guessable passwords.	Moderate	High	Moderate	2
2	Unpatched vulnerabilities in the GenAI codebase.	High	High	High	1
3	Poorly secured API endpoints prone to injection attacks.	High	Moderate	Moderate	2

4	Misconfiguration during server setup.	Moderate	High	Moderate	2
5	Lack of timely updates and patches.	Moderate	High	Moderate	2
6	Default credentials not changed, leading to unauthorized access.	High	High	High	1
7	Lack of encryption, leading to data exposure during extortion.	Moderate	High	Moderate	2
8	Single point of failure in server hardware.	Low	High	Moderate	3
9	Insufficient encryption on wireless communication.	Moderate	Moderate	Moderate	3
10	Bugs and code performance issues leading to system crashes.	High	High	High	1
11	Lack of physical security, leading to tampering.	Moderate	High	Moderate	2
12	Lack of disaster recovery plans for natural disasters.	Low	High	Moderate	3
13	Weak access controls leading to unauthorized data access.	High	High	High	1
14	Insufficient monitoring and auditing of employee activities.	Moderate	High	Moderate	2
15	Compromised third-party software or components used within the application.	Moderate	High	Moderate	2
16	Insufficient rate limiting and authentication mechanisms on APIs.	High	Moderate	High	2
17	Weak password policies leading to credential theft.	High	High	High	1

18	Improper configuration of cloud services, leading to exposure of sensitive data.	Moderate	High	Moderate	2
19	Unpatched zero-day vulnerabilities exploited by attackers.	Moderate	High	Moderate	2
20	Lack of adherence to data protection and cybersecurity regulations, leading to legal and financial penalties.	Moderate	High	Moderate	2
21	Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs.	High	High	High	1
22	Attackers reverse-engineering the model to infer sensitive information from the training data.	Moderate	High	Moderate	2
23	Small, crafted inputs causing the GenAI model to make incorrect or harmful predictions.	High	High	High	1
24	Performance degradation over time as the model encounters data differing from its training set.	Moderate	Moderate	Moderate	3
25	Generation of nonsensical or incorrect outputs that appear plausible.	Moderate	Moderate	Moderate	3
26	Inherited biases from training data leading to unfair or discriminatory outputs.	Moderate	High	Moderate	3
27	Difficulty in understanding and interpreting model outputs,	Moderate	Moderate	Moderate	3

	obscuring errors and reducing trust.				
28	Inadequate or biased training data leading to flawed models and erroneous outputs.	Moderate	High	Moderate	2
29	Unintentional plagiarism or copyright infringement due to the replication of protected content.	Moderate	High	Moderate	2
30	The model performs well on training data but poorly on new, unseen data.	Moderate	Moderate	Moderate	3
31	GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks.	High	High	High	1
32	The application inadvertently exposing sensitive personal or organizational data through its outputs.	Moderate	High	Moderate	2
33	Attackers manipulating the GenAI's algorithm to produce biased or harmful outputs.	High	High	High	1
34	Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs.	High	High	High	1
35	GenAI producing outputs that violate ethical standards, such as generating offensive content.	Moderate	High	Moderate	2

Table-9 Vulnerability occurrence and Risk Assessment Table

2.3.7 Risk Mitigation Strategies:

To address the identified vulnerabilities and associated risks effectively, a comprehensive risk mitigation strategy is developed. According to the NIST Cybersecurity Framework (CSF) 2.0 To organize vulnerabilities and corresponding mitigation strategies, each vulnerability is categorized by NIST core functions (Identify, Protect, Detect, Respond, Recover and Govern) with clear mappings to the respective mitigation strategies (Dimakopoulou and Rantos 2024). This strategy includes preventative measures, detection mechanisms, and response plans tailored to the specific vulnerabilities and their risk rankings (Katsumata et Al. 2010), (Blank and Gallagher 2012). Robust mitigation strategies categorized by NIST CSF category and sub-category risk rank are as Follows:

2.3.7.1 High-Risk Vulnerabilities

1. **Unpatched Vulnerabilities (Vul No. 2):**
    - **Category:** PROTECT
    - **Subcategory:** Platform Security (PR.PS)
    - **Mitigation:** Align with PR.PS-02 (Software is maintained, replaced, and removed commensurate with risk).

    Implement a robust patch management process. Regularly update and patch all software components, including third-party libraries. Conduct periodic security audits and vulnerability scans to identify and address potential issues. Default.

3. **Credentials Not Changed (Vul No. 6):**
    - **Category:** PROTECT
    - **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
    - **Mitigation:** Align with PR. AA-01 (Identities and credentials for authorized users, services, and hardware are managed by the organization).

      Enforce strong password policies and ensure that default credentials are changed during initial setup. Use multi-factor authentication (MFA) to enhance access security.

4. **Bugs and Code Performance Issues (Vul No. 10):**
    - **Category:** PROTECT
    - **Subcategory:** Platform Security (PR.PS)
    - **Mitigation:** Align with PR.PS-06 (Secure software development practices are integrated, and their performance is monitored throughout the software development life cycle).

Establish a rigorous code review and testing process. Implement automated testing tools to detect bugs and performance issues early in the development cycle. Conduct regular performance evaluations and stress testing.

5. **Weak Access Controls (Vul No. 13):**
    - **Category:** PROTECT
    - **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
    - **Mitigation:** Align with PR. AA-05 (Access permissions, entitlements, and authorizations are defined in a policy, managed, enforced, and reviewed, and incorporate the principles of least privilege and separation of duties).

      Implement granular access controls and enforce the principle of least privilege. Regularly review and update access permissions. Use MFA and strong authentication mechanisms for sensitive systems.

6. **Weak Password Policies (Vul No. 17):**
    - **Category:** PROTECT
    - **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
    - **Mitigation:** Align with PR. AA-01 (Identities and credentials for authorized users, services, and hardware are managed by the organization).

      Develop and enforce strong password policies, including complexity requirements and regular password changes. Implement MFA for critical systems and accounts.

7. **Malicious Data in Training Datasets (Vul No. 21):**
    - **Category:** IDENTIFY
    - **Subcategory:** Risk Assessment (ID.RA)
    - **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

      Implement data validation and cleansing procedures to detect and remove corrupted or malicious data. Use secure data sources and regularly review training data for integrity.

8. **Reverse-Engineering of the Model (Vul No. 22):**
    - **Category:** PROTECT

- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-05 (Installation and execution of unauthorized software are prevented).

  Employ model protection techniques, such as encryption and obfuscation, to safeguard intellectual property. Regularly review and update security measures around model deployment

9. **Small, Crafted Inputs Causing Harmful Predictions (Vul No. 23):**
   - **Category:** DETECT
   - **Subcategory:** Continuous Monitoring (DE.CM)
   - **Mitigation:** Align with DE.CM-09 (Computing hardware and software, runtime environments, and their data are monitored to find potentially adverse events).

     Implement robust input validation and sanitization processes. Use anomaly detection mechanisms to identify and mitigate malicious inputs.
10. **GenAI Used for Social Engineering Attacks (Vul No. 31):**
    - **Category:** DETECT
    - **Subcategory:** Adverse Event Analysis (DE.AE)
    - **Mitigation:** Align with DE.AE-02 (Potentially adverse events are analyzed to better understand associated activities).

      Monitor and restrict the use of GenAI for generating sensitive or potentially harmful content. Implement detection mechanisms to identify misuse of the model.
11. **Algorithmic Manipulation (Vul No. 33):**
    - **Category:** PROTECT
    - **Subcategory:** Platform Security (PR.PS)
    - **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

      Apply security controls to prevent unauthorized access and modifications to the GenAI algorithm. Conduct regular audits to detect and address potential manipulations.
12. **Unauthorized Access to GenAI Application (Vul No. 34):**
    - **Category:** PROTECT

- **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
- **Mitigation:** Align with PR. AA-06 (Physical access to assets is managed, monitored, and enforced commensurate with risk).

  Strengthen access controls and implement logging and monitoring to detect unauthorized access attempts. Use encryption and secure authentication mechanisms.

2.3.7.2 Moderate-Risk Vulnerabilities

13. **Weak Authentication (Vul No. 1):**
    - **Category:** PROTECT
    - **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
    - **Mitigation:** Align with PR. AA-02 (Identities are proofed and bound to credentials based on the context of interactions).

      Strengthen authentication mechanisms by enforcing strong password policies and implementing MFA. Regularly review and update authentication methods.
14. **Poorly Secured API Endpoints (Vul No. 3):**
    - **Category:** PROTECT
    - **Subcategory:** Platform Security (PR.PS)
    - **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

      Secure API endpoints with proper authentication and authorization mechanisms. Implement rate limiting and input validation to protect against injection attacks.
15. **Lack of Timely Updates and Patches (Vul No. 5):**
    - **Category:** PROTECT
    - **Subcategory:** Platform Security (PR.PS)
    - **Mitigation:** Align with PR.PS-02 (Software is maintained, replaced, and removed commensurate with risk).
    - Develop a patch management strategy to ensure timely application of security updates. Monitor for new vulnerabilities and apply patches as needed (Huang et al. 2024).

16. **Lack of Encryption for Data Exposure (Vul No. 7):**
    - **Category:** PROTECT
    - **Subcategory:** Data Security (PR.DS)
    - **Mitigation:** Align with PR.DS-01 (The confidentiality, integrity, and availability of data-at-rest are protected).
      Implement strong encryption for data at rest and in transit. Regularly review encryption practices to ensure they meet current security standards.
17. **Insufficient Monitoring and Auditing (Vul No. 14):**
    - **Category:** DETECT
    - **Subcategory:** Continuous Monitoring (DE.CM)
    - **Mitigation:** Align with DE.CM-01 (Networks and network services are monitored to find potentially adverse events).
      Enhance monitoring and auditing processes to detect and respond to suspicious activities. Implement comprehensive logging and regular reviews of access and activity logs.
18. **Compromised Third-Party Components (Vul No. 15):**
    - **Category:** GOVERN
    - **Subcategory:** Cybersecurity Supply Chain Risk Management (GV.SC)
    - **Mitigation:** Align with GV.SC-07 (The risks posed by a supplier, their products and services, and other third parties are understood, recorded, prioritized, assessed, responded to, and monitored over the course of the relationship).
      Evaluate and vet third-party software and components for security. Regularly update and patch third-party components and use trusted sources.
19. **Insufficient Rate Limiting on APIs (Vul No. 16):**
    - **Category:** PROTECT
    - **Subcategory:** Platform Security (PR.PS)
    - **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).
      Implement rate limiting and API gateway controls to manage and restrict API traffic. Monitor API usage for unusual patterns (Huang et al. 2024).
20. **Improper Configuration of Cloud Services (Vul No. 18):**
    - **Category:** PROTECT
    - **Subcategory:** Platform Security (PR.PS)

- **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

  Regularly review and audit cloud service configurations to ensure compliance with best practices. Implement security controls for data protection and access management.

21. **Unpatched Zero-Day Vulnerabilities (Vul No. 19):**
    - **Category:** IDENTIFY
    - **Subcategory:** Risk Assessment (ID.RA)
    - **Mitigation:** Align with ID.RA-02 (Cyber threat intelligence is received from information sharing forums and sources).

      Stay informed about emerging vulnerabilities and apply updates as soon as patches become available. Use threat intelligence to anticipate and prepare for zero-day threats (Krishnamurthy 2023).

22. **Lack of Adherence to Regulations (Vul No. 20):**
    - **Category:** GOVERN
    - **Subcategory:** Organizational Context (GV.OC)
    - **Mitigation:** Align with GV.OC-03 (Legal, regulatory, and contractual requirements regarding cybersecurity — including privacy and civil liberties obligations — are understood and managed).

      Ensure compliance with data protection and cybersecurity regulations through regular audits and updates to policies and procedures. Provide training to employees on regulatory requirements.

23. **Inherited Biases in Training Data (Vul No. 26):**
    - **Category:** IDENTIFY
    - **Subcategory:** Risk Assessment (ID.RA)
    - **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).
    - Regularly review and audit training data for biases. Implement techniques to detect and mitigate bias in model training.

24. **Inadequate or Biased Training Data (Vul No. 28):**
    - **Category:** IDENTIFY
    - **Subcategory:** Risk Assessment (ID.RA)

- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).
- Use diverse and representative datasets for training. Regularly review data quality and make necessary adjustments to improve model accuracy and fairness (Krishnamurthy 2023).

25. **Unintentional Plagiarism (Vul No. 29):**
    - **Category:** IDENTIFY
    - **Subcategory:** Risk Assessment (ID.RA)
    - **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

      Implement measures to detect and avoid replication of protected content. Use plagiarism detection tools and adhere to copyright laws.
26. **Privacy Invasion Through Outputs (Vul No. 32):**
    - **Category:** PROTECT
    - **Subcategory:** Data Security (PR.DS)
    - **Mitigation:** Align with PR.DS-01 (The confidentiality, integrity, and availability of data-at-rest are protected).

      Implement data anonymization techniques and review outputs to ensure they do not expose sensitive information.
27. **Ethical Violations (Vul No. 35):**
    - **Category:** GOVERN
    - **Subcategory:** Organizational Context (GV. OC)
    - **Mitigation:** Align with GV. OC-01 (The organizational mission is understood and informs cybersecurity risk management).

      Establish ethical guidelines for GenAI usage and ensure compliance through regular reviews and audits. Implement mechanisms to detect and address unethical outputs.

2.3.7.3 Low-Risk Vulnerabilities

28. **Single Point of Failure in Server Hardware (Vul No. 8):**
    - **Category:** PROTECT
    - **Subcategory:** Technology Infrastructure Resilience (PR.IR)

- **Mitigation:** Align with PR.IR-03 (Mechanisms are implemented to achieve resilience requirements in normal and adverse situations).
  Implement redundancy and failover solutions to mitigate the impact of hardware failures. Regularly test disaster recovery plans.

29. **Insufficient Encryption on Wireless Communication (Vul No. 9):**
    - **Category:** PROTECT
    - **Subcategory:** Data Security (PR.DS)
    - **Mitigation:** Align with PR.DS-02 (The confidentiality, integrity, and availability of data-in-transit are protected).
    - Use strong encryption protocols for wireless communications. Regularly review and update encryption practices.
30. **Lack of Disaster Recovery Plans (Vul No. 12):**
    - **Category:** RESPOND
    - **Subcategory:** Incident Recovery Plan Execution (RC.RP)
    - **Mitigation:** Align with RC.RP-01 (The recovery portion of the incident response plan is executed once initiated from the incident response process).
      Develop and test comprehensive disaster recovery plans. Ensure plans address various scenarios and include regular updates.
31. **Performance Degradation Over Time (Vul No. 24):**
    - **Category:** IDENTIFY
    - **Subcategory:** Risk Assessment (ID.RA)
    - **Mitigation:** Align with ID.RA-04 (Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded).
    - Implement monitoring tools to track model performance and conduct regular evaluations. Update models as necessary to address performance issues.
32. **Generation of Nonsensical Outputs (Vul No. 25):**
    - **Category:** IDENTIFY
    - **Subcategory:** Risk Assessment (ID.RA)
    - **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).
      Continuously review and refine the model to improve output quality. Implement feedback mechanisms to detect and address nonsensical outputs.

33. **Difficulty in Interpreting Model Outputs (Vul No. 27):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

  Enhance model explain ability and provide clear documentation on output interpretation. Use visualization tools to aid in understanding outputs.

34. **The Model Performs Poorly on New Data (Vul No. 30):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-04 (Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded).

  Implement continuous learning and adaptation strategies to improve model performance on new data. Regularly update the training data to reflect changing patterns.

2.4 Collaborative Training and Awareness Program

Organizations are gradually enhancing the qualitative aspects of cybersecurity through the inclusion of artificial intelligence in their organizational security models, therefore highlighting the necessity of a versatile training program that may meet the needs of technical as well as non-technical personnel. The reliability of such developed instruments in protecting digital assets mainly depends on the level of understanding of the entire personnel.

2.4.1 Reason for a Thorough Training

In the sphere of cybersecurity, AI-driven solutions radically change the look of threat identification and risk management processes. These tools are very robust in terms of functionality and efficiency for threat detection and risk management but their effectiveness solely and objectively depends on the capabilities of the respective users. Consequently, there is a necessity of implementing a wide training program, allowing considering all potential and actual employees as staff, which must be ready to be involved into security as contributors.

2.4.2 Training Objectives and Structure

The training program must address the distinct needs of both technical and non-technical staff:The training program must address the distinct needs of both technical and non-technical staff:

1.Technical Staff: This group, which comprises IT specialists as well as cybersecurity, must be trained in detail about all the procurement AI tools utilized in identifying website vulnerabilities. This will entail the training covering such topics as the employment and setting up of such tools, comprehensive examination of the AI derived reports and methods of solving problems. Skills-based sessions will prove vital: pragmatic and largely applicable, it will be the type of training where the participants will be exposed to real life situations and practice responses to various security considerations. Consequently, there is always a focus on training so that security may adapt to new threats and the developments in AI.

2. Non-Technical Staff: Even workers that do not engage in technical roles therefore have a role to play in the organizations cybersecurity framework. Their training comprises of a basic training that entails an understanding of the websites' weaknesses and how AI can help in mitigating the same. Importance is placed on practicing high levels of security at personal interactions with computers, the internet and other gadgets such as playing safe on the internet, or identify phishing scams. The training will also teach communication of potential security issues to technical people on how to communicate them better. Regardless of this, non-technical content that will be captivating to the eyes and easy to grasp will meet this objective

of reaching out to staff.

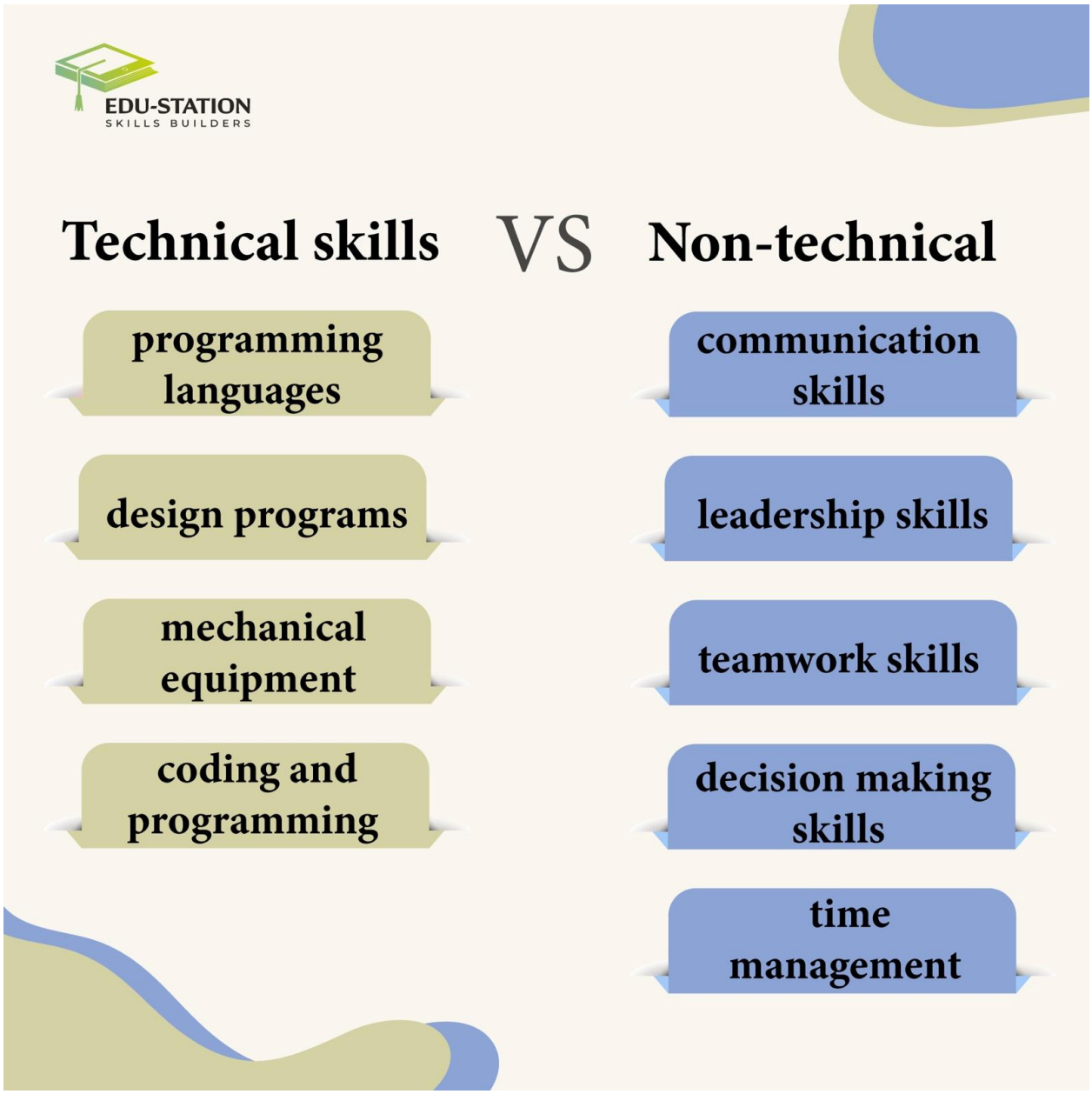


Figure-10 :Technical Skills and Non-Technical Skills

2.4.3 Training Program for Technical and Non-Technical Staff: AI Tools for Website Vulnerability Detection

For both technical as well as non-technical employees to be capable of properly implementing and supporting artificial intelligence tools especially those that help in reconnaissance on sites like the Water tunnel Carwash site, the practical knowledge need to be imparted. This training is focused to improve the security status of websites and to avoid disruption which can hinder the performance and productivity of employees.

1.Technical Staff:

- **Proficiency in AI Tools:**

  The training will include an overview of how to install and set up AI tools and will provide the technical staff with the understanding of the procedures of using these tools in both cloud services and on local hardware. They will be able to install the tools to meet system requirements, solve installation problems and structure the tools to fit the Water tunnel Car wash's security situation. After that the training will include the training on Deployment and Integration Subjects to elaborate on progressive and big data deployment, APIs integration and data flow for improving cybersecurity systems and handling changes in security specifications.

  Regarding the staff, the probably key impact would be the development of the comparative analysis and usage of the findings in the reports generated through the application of AI tools. They will also know how to analyse various forms of vulnerability reports, associate it with the threat models hence come up with better security decisions. This module will be geared toward setting specific recommendations of how the artificial intelligence tools must be used as well as how they can be fine-tuned by the staff to deliver better results.

- **Advanced Troubleshooting Skills:** In the training program, act as a priority in the training program will be higher level troubleshooting skills to track and manage any issues that may arise with AI tools being utilized in the identification of vulnerabilities at the site of Water tunnel Carwash. Employees will be able to analyse complex problems related to software, its configurations, and with interfaces between different applications. They will also learn methods of diagnostics, for example, understanding error messages and logs to identify the sources of issues. The training will also include systematic approach to problem solving that will allow the staff determine the nature of problems, trial possible solutions, and verify fixes systematically.

  Also, staff will learn how to apply the solution, such as patches for a host of problems, and adjustments to configurations, without significantly disrupting others and how to quickly get back on-line. Documentation will be highly appreciated focusing on the descriptions of troubleshooting processes and solutions with the aim to reach out to relevant information that will be highly useful in the future working processes and will also encourage the employees to share their problems and ideas. Finally, perpetuity of

improvement will be inevitable, and the staff will be urged to analyse previous issues, look for patterns, and contribute their inputs for the optimization of both, how-diagnostics, and AI instruments. Such an approach guarantees that the staff needed to adjust the functionality of the applied AI solutions and protect the websites' cybersecurity are prepared correctly (Frederiksen and White, 1998).

- **Integration with Security Systems:** On the training program, staff will dedicate their time and efforts to learning how to incorporate AI tools in existing frameworks so that the tools are complementary to other security mechanisms. This portion of the training is to start by presenting information about existing security systems such as firewalls, IDSs and other security technologies that are employed. Staff will be trained on how to incorporate AI tools with these systems where data will be freely passed between them and the tools put in the proper security context as build in the AI systems.

  To this end, the training will address APIs, and data sharing standards through which the AI applications talk to the currently installed security systems. The staff members will get practical experience in tuning these integrations so that they can harmonize and run at the best way. Moreover, they will also be trained to overcome integration issues which may arise like conflicts with other security systems or modifying the AI tools to full certain security objectives.

  Furthermore, by training staff in the identification of these integrations, they would also be concurrently trained on how to assess its efficacy in boosting the security profile of the system while not compromising by bringing about fresh points of failure. This encompasses; parameterization and performance comparison, testing, and some configurations. In this way, staff will take procedures to make sure that incorporating AI tools as enhances will not disrupt the effectiveness of existing security measures but complements them and enhance the already existing layers of security (Boiko and Shendryk, 2017).

2. Non-Technical Staff

- **Basic Awareness of Website Vulnerabilities:** During the training non-technical employees will learn about the more frequently used methods of web-site vulnerabilities, and their relation to AI tools. It provides a verbally descriptive, easily

understandable picture of threats that confront web sites to the staff of a company for a specialized security understanding excluding professional computer knowing (Appiah et al., 2017).

- **Understanding Common Vulnerabilities**: The common website threats that the staff will be familiar with include cross site scripting (XSS), SQL injection, cross site request forgery (CSRF) and security misconfigurations. The training will elaborate the fact that how these vulnerabilities can be used by attackers to put website security at risk and the impact for the organization. When staff is aware of these threats, the indicators of possible security problems will be more easily identifiable, and the threats will be appreciated for the significance they have in the context of security (Rajapaksha et al., 2022).
- **Role of AI Tools:** The training will then proceed to demonstrate how vulnerable uses artificial intelligence in the identification as well as mitigation of these risks. Employees will become familiar with the fact that AI can be utilized for vulnerability scanning, threat detection in real-time mode as well as predictive analysis. They will learn how these tools aids in discovering exploits which could be hard to find manually, and how they come in handy in a proactive defence. This will cover such areas as; how AI technologies work in the identification of trends from large datasets, how they learn and how they assist in improving website security.
- **Practical Applications**: To help staff better identify with the technical presentations as well as support this workshop, staffs will be demonstrated with an example of a report format that vulnerability AI tools produce and the sort of data the report delivers. They will also learn the most basic level of reporting interpretation in that they will be able to know about the identified vulnerabilities and the most appropriate recommendations. It will help the staff in being able to contribute towards the process of handling the vulnerabilities and helping the technical team in their work (Ahmad, N. and Alsmadi, I., 2021)

- **Security Best Practices:** First, during the training program, the staff will be provided with a comprehensive knowledge of measures that should be adopted to have a secure online environment and adequate measures for the identification of security loopholes. Employees will first begin with the basics, including setting up good and different

password and the practice of MFA for accounts. These will be equipped with various advanced procedures in handling the data to make sure that sensitive data is protected using encryption as well as managing the data in a right manner. The program will also focus on the indication and reporting of security threats, where the staff will be trained on how to notice signs of security threats such as odd symptoms and communications pertaining to the systems then raise the alarm. Also, safe browsing and emails: Various tips on how to avoid getting to the wrong website and handling email attachments will be addressed among the employees. The awareness on the need for frequent checks and upgrading of the software to avoid the shortcomings arising from outdated systems shall be brought out. Last of all, it will be required that the staff should continually refresh his or her knowledge on the cybersecurity trends and challenges. This is a holistic approach for ensuring that staff are well equipped to protect the online integrity and leverage their efforts to the organization's cybersecurity strategies.

- **Effective Reporting and Communication:** In the training program, non-technical staff will be guided on effective reporting and communication practices to support vulnerability management. Staff will learn to recognize and document anomalies or suspicious activities, such as unusual system behaviour or irregular access patterns, and promptly report these issues using internal reporting tools. The training will cover how to provide clear, detailed descriptions of the observed problems to ensure accurate and efficient investigation by technical teams. Additionally, staff will be trained in basic technical terminology to facilitate meaningful communication with technical teams, helping them provide necessary context and background information. They will also be taught how to collaborate with technical teams during incident resolution, including maintaining open communication, providing updates, and following instructions. Emphasis will be placed on documenting the reporting process and offering feedback to improve procedures. This comprehensive approach ensures that non-technical staff effectively contribute to identifying and addressing security issues, enhancing the organization’s overall cybersecurity efforts.

2.4.4 Training Module for Technical Staff

Module 1: Introduction to AI Tools for Vulnerability Detection

This module aims designed to acquaint the technical staff with the tools in AI-based vulnerability detection in the newly developed Water tunnel Carwash site. Introducing the AI tools that shall be used in the module, the module's content shall include an analysis of how the tools work, their key features and how they can help automate the process of vulnerability scanning and threat detection as well as generating real-time reports. It will then move to discuss how these tools are constructed using AI and machine learning including the notion of supervised and unsupervised learning, neural networks as well as how these algorithms employ pattern matching to look out for weaknesses. Further, the staff will get to know how those AI tools interact with the existing security systems, how to set up, integrate such tools with firewalls, with IDS and with any other security systems? The training will involve presentation supported by interactive features and demonstrations thus providing the staff with practical view of the tools as well as engaging them in questions and answers section that offers them a better understanding of the practical applicability of the tools.

Module 2: Hands-On Configuration and Deployment

This module is designed to provide technical staff with hands-on experience in setting up and implementing AI tools for vulnerability detection. The module will begin with detailed instructions on installation procedures and configuration settings, covering system requirements, software prerequisites, and key configuration options such as scan parameters and user permissions. Staff will be guided through best practices for deployment, learning strategies to optimize tool performance, ensure compatibility with existing systems, and avoid common pitfalls such as integration issues and configuration errors. The practical component will include real-world simulations, allowing staff to apply their knowledge in a controlled environment and practice deploying the AI tools while addressing simulated challenges. These lab sessions will be conducted with step-by-step guidance from instructors, ensuring that staff gain confidence and develop problem-solving skills essential for successful tool deployment. This hands-on approach will equip staff with the practical experience needed to effectively configure and deploy AI tools within the organization's security framework.

Module 3: Advanced Analysis and Troubleshooting

This module concerns primarily with enhancement of technical staff's competence in assessing and enhancing efficiency of AI application for vulnerability identification. The proposed module is to improve staff's willingness to read AI-derived vulnerability reports, the ability to comprehend the importance of certain observations and to make the right decision regarding

the remediation. This will also tackle the issues related to false positives and negative, the proposed training will allow the staff to recognize and minimize errors between the results of the AI tools, improve detection and accuracy in the results. Moreover, the recognitions of technical problems, interpretation of error logs, identification of troubleshooting processes and practice of maintenance will also be within the aspects to be discussed in this module of the study. Pandemic care and presented as a workshop, which includes using case and exercise in the form of 'troubleshooting', this training would offer practical practice on how to handle with actual situation. Using organization's reports staff will exercise to go through, seeing how to work with mistakes, and how to solve certain technical difficulties, so they will build confidence and capability in terms of working with more complex level of collecting data for AI and dealing with arising glitches.

Module 4: Ongoing Learning and Tool Updates

concerned in providing information's about the newly invented tools in the field of artificial intelligence and the security risks involved in it to technical staffs. This module should help the staff continue to abreast with the new changes and updates made on the AI tools to make the most out of the newest features and changes of the tool. Furthermore, the module will identify new security threats with relation to the Water tunnel Carwash site; this part of the training will ensure that the site's staff is informed of new risks and ways in which a site may be compromised. Presented via webinars of briefings organized by the development team, the module will provide updates and enable dialogue on the most recent events. This approach ensures that staff are continually learning and enhancing their knowledge about cybersecurity which will help them in the management as well as optimisation of AI tools to meet the challenges that may occurred in the future.

Non-Technical Staff

Module 1: Basics of Website Vulnerabilities and AI Tools

This module designed to enlighten nontechnical workers about website vulnerabilities and the use of artificial intelligence tools in solving such vulnerabilities. Staff will be presented to the well-known website weaknesses, including SQL injection, cross-site scripting (XSS), and cross-site request forgery (CSRF) and their consequences for the organization's security. This way the staff would know how those vulnerabilities can be utilised and what should be expected, which will help them appreciate the concept of cybersecurity. The training will also describe how AI solutions work to identify and mitigate such threats as well as how they work

best when it comes to scanning for the problems, pattern analysis, and report generation. This will assist the staff to understand how AI tools can be of value in the management of vulnerabilities and the general support for the company's security plan. Further, the module will focus on increased awareness regarding the threats, habits, as well as reporting mechanisms when it comes to cyber threats. To increase understanding of the materials delivered, an interactive presentation of the module in the form of diagrams and videos will be used to capture staff's attention. This approach will make sure that all employees especially those non-technical ones are aware of Website vulnerabilities and AI Tools for security (Ranimäki, 2023).

Module 2: Security Best Practices and AI Awareness

Security Best Practices and AI Awareness is aimed to provide non-IT employees with valuable information on how to prevent insecurity threats and learn more about AI solutions applied in the organization's security management. In a learning module for cybersecurity, students shall be trained on issues to do with protection on the internet including password creation, phishing, and connection security. It will focus on such approaches that the staff at an organization can use to protect such accounts and information. Also, it will be important to teach when and how to report certain possible security threats such as emails received, links to click, and system responses. To such cases, staff will be taught on how to handle and report such incidences to the IT or the security department appropriately. The module will also be talk about different aspects of use of AI in security processes, where these tools will explain how they assist in the process of identifying threats, analysing risks, and making decision. Since the described module is going to be provided in the form of a workshop involving role play and real-life case scenarios, the staff will be able to replay actual security scenarios learn how to apply best practices and find out the role of AI tools in the risk management process. This way and style will guarantee that even the staffs, who are non-technical, will be able to continue with the process.

This is because the non-technical will be put through drill in ensuring they can work and handling matters of security should anything go wrong (A., Myneni. ,et. al,2019).

Module 3: Reporting and Collaboration

This training module helps non-IT personnel acquire knowledge in the way and manner they will report security incidents as well as how to interface with IT personnel. It includes process of recording and reporting the issues identified by the artificial intelligence systems, improving

the ways of explaining problems to stakeholders, and recognizing the specific terms used by technical experts. The module also looks at teamwork mechanisms to address risks with specific regard to the non-IT staff. When students take mock sessions, exposure and understandings are achieved by the staff so that they can meaningfully proceed with the security measures and cooperate properly with the technical staff.

Module 4: Continuous Awareness and Engagement:

This module will help to support the engagement of non-technical staff of the company through consistent updates on the AI tools and security solution. Users and staff will be informed about new features, results, and security updates, as well as be occasionally reminded of correct password usage, or the risks of phishing attacks. Semi-monthly newsletters and brief workshops will provide brief overviews, including staff-members' tasks regarding security and training them to be ready to ensure cybersecurity of the organization.

Module 5: Assessment and Feedback

This module is designed to assess the effectiveness of the training program and gather insights for improvement. It includes online quizzes to evaluate staff's understanding of AI tools, security practices, and reporting procedures. Additionally, feedback forms will collect staff opinions on the training's relevance, clarity, and practical application. By combining assessments with feedback, the module ensures the training remains effective and evolves to meet staff needs, enhancing overall security practices within the organization.

2.4.5 Implement and Delivery

Scheduling

The implementation of the training program will be carefully scheduled to maximize effectiveness and accommodate the different needs of technical and non-technical staff.

For technical staff the training will incorporate a time frame where the flow of the training will be divided into different stages that will follow a module approach to the training. Each of them will be reconsidered as a part of the consecutive tightly interconnected modules, which will provide technical staff enough time for the practice and elimination of possible troubles. This approach will ensure that staff get to be quite acquainted with the AI tools and how they can be deployed for vulnerability detection, the installation, configuration as well as optimisation of the same. The timings are to be divided between both theoretical

and practical sessions done in Laboratory The course will also include lab days which are the proactive days to turn the concept into practically applicable tools.

The training for the non-technical employees will be conducted in several sessions which are broken down to enable easy assimilation. The intent of these sessions will be to go over important ideas and helpful information and teaching therapy which will not overwhelm the staff. Blended lessons will be created to enhance the knowledge retention of the staff and to provide information concerning changes or novelties in the sphere of security as well as AI. This will all be made possible to make certain that all other employees who are not technical will not have to lose touch with the subject matter as well as continually get a better understanding of their roles in relation to security.

Altogether, the scheduling will be made on an individual basis considering all the peculiarities of groups and providing the staff members with the necessary training at the proper rate to achieve proper understanding and implementing of the security practices.

Delivery Methods

To effectively deliver the training program to both technical and non-technical staff, a variety of tailored methods will be used to address their distinct needs and learning preferences.

- **For technical staff**: Targets of training are in-person workshops for true understanding of the AI concepts through interaction, practical AI labs for mastering necessary tools, and online webinars for the more convenient way to enhance skills and updates. Demonstration and discussion-based workshops will be conducted as will laboratory-based sessions involving practical examples and webinars will be used to provide updates on new approaches and detailed methods.
- **For non-technical staff**: In this case, the training will comprise of online Mear modules and web-based workshops. Pre-recorded webinars and e-learning modules will offer a well-defined systematic course delivery method where staff can freely access pertinent material at will. These modules will incorporate some form of media to pass knowledge in form of videos, infographics, and quizzes among others. Periodically, useful workshops will be held as a way of reminding the staff on the important concepts to observe as well as giving them practice time in reporting and collaborating, but in a safe environment. These workshops are also going to be based on role-play and scenarios so that students do not feel they are in passive learning environment and thus

would be more receptive to the content delivered in the workshops. Also, reports with related information will be submitted monthly either in the form of newsletters or brief briefings so that the non-technical staff can be updated with the current trends on security practices and the application of Artificial Intelligence. Maintaining regular meetings like these will also assist in cementing the learning when it comes to security issues and the staff will also not lose touch on the activities in the concerned department.

Combining these methods of delivery, the training program will cover all requirements of technical and non-technical employees to guarantee that all human resources of the organization are ready to contribute to cybersecurity efforts (Furnell, S. and Clarke, N., 2012).

2.4.6 Support and Resources

- **Technical Staff**: Technical staff that will manage the application will be fully supported with manual, setup and technical assistance documentation. Online forums will enable people to share experiences with each other while getting a direct access to the developers will ensure that customers get proper assistance. These resources must enable the staff to have proper understanding of how to properly deploy AI tools in detecting vulnerabilities.
- **Non-Technical Staff:** Some of the features that will be made available to the non-technical will be a guideline that will enable them to identify the vulnerability and the procedure to follow when reporting the issues, frequently asked questions and answers, and a support help desk where additional assistance will be given. Then these resources will assist them in applying security best practices as well as managing security tasks proactively.

2.4.7 Evaluation and Continuous Improvement

Assessment

To measure the effectiveness of the training program and ensure that both technical and non-technical staff have acquired the necessary skills and knowledge, targeted assessments will be implemented.

- **Technical Staff**: Assessments for technical staff will include practical evaluations, such as configuring AI tools and troubleshooting, to test real-world application skills. Theoretical assessments will involve written tests or quizzes on key concepts and protocols. This combination will ensure a thorough evaluation of both practical and theoretical knowledge.
- **Non-Technical Staff**: Non-technical employees will participate in knowledge-based security questions, and questions which would test how they would implement them practically. Such assessments will help confirm that security is communicated to the staff and that they are ready to encounter challenges in a real-life environment while working together with tech teams.

Feedback

The training program will also comprise surveys administered throughout the year to participants that are both technical and non-technical to observe the extent to which the program is fulfilling the participants' needs and make necessary alterations if necessary.

- **For technical staff:** Students' feedback will be collected by administering surveys and evaluations of the course content in the format of neatly arranged written materials, relevance of in-class exercises, effectiveness of laboratories. The technical staff will also complete the survey where they will estimate the difficulty of the tasks, the available support and their confidence level with the AI tools; these will also need to be reviewed and improved to come up with the most efficient training program possible.
- **For non-technical staff:** Data will be obtained by using short questionnaires other than formal assessments to assess the understanding, and relevance and preparedness of the staff concerning security measures. Employees who are not technical will also contribute to the instructional content and the dynamic components. In the process of implementing the training program, feedback will be collected at different intervals to ensure the continued effectiveness of the program and capture changes staff might require (Schmoker, 1999).

Continuous Learning

Continuous Learning is a vital element of the training program, designed to ensure that staff remain current with evolving developments and best practices in cybersecurity. This ongoing educational approach helps maintain high standards of security and adaptability within the organization.

Ongoing-Updates

The program is expected to disseminate to clients on a frequent basis the latest information on the state of artificial intelligence tools, features, and trends; the latest on the security of the system. Such updates will assist the staff in being informed and or employing new tools and techniques in vulnerability detection.

Refresher-Courses

Periodically offered refresher courses will revisit core topics like security best practices and AI tool functionalities. These concise, focused sessions aim to refresh staff's knowledge, address any gaps, and reinforce practical application without overwhelming them.

Format-Delivery

For repetition and continued education there will be webinars, on-line courses and classes conducted in person intermittently to cover different learning styles. By doing so, the staff is aware of the current trends thus improving proficiency and facilitating utilisation of the AI tools.

2.4.7 Raising Awareness about the Capabilities, Limitations, and Best Practices of GenAI

Ensuring the safe and effective use of Generative AI (GenAI) technologies requires comprehensive awareness and understanding among all stakeholders. Here's a detailed approach to raising this awareness:

1. Educate on Capabilities

Understanding Capabilities

- **Training Sessions:** There is, therefore, the need to conduct training sessions on Generative AI (GenAI) to optimize usage. These sessions are to focus on basic features of GenAI including content creation, routine work handling, and predictive capabilities. It will also show how report writing can be done automatically, how graphic designs for marketing can be produced, voice-overs created, and trends predicted using GenAI. Training will be needs based value-added sessions and use of other software and will incorporate exercise and group discussions. Subsequent material and further coaching will maintain staff's awareness of further developments in GenAI. It makes certain all the employees are ready to utilize GenAI to its productive best by removing barriers.

- **Workshops and Demonstrations:** This shows that training sessions will involve the stakeholders in the use of Generative AI tools so that they are conversant with real time text image and audio generation as well as analysis of data. Such live coherent sessions will also expose how GenAI can develop content, predict trends, and improve organisational services. GenAI is designed to be implemented with ease, and via exercises and question and answer, sessions, stakeholders will discover how best to adopt the tool to enhance productivity.

2. Demonstrations and Case Studies

- **Case Studies:** Sharing of research, comparing successes of using Generative AI (GenAI) in industries that are like the targeted industry provides great information to the stakeholders. For instance, a retail case study on GenAI can demonstrate how the technology aided in improving the aspects of personalization and boosting the sales. For example, a case study in the field of finance could show how GenAI optimised the analysis of data on risks and their management, thus enhancing the quality of the decision made and saving money. In a case of healthcare, a case could focus on how GenAI automated much of the bureaucratic work in hospitals making care even better for the patients and lighter for the employees. It is expected that each case report will highlight the issues encountered, the GenAI solutions used, and the results obtained to provide practical implications for improvement for GenAI to be used by stakeholders.
- **Success Stories:** Another effective way is to tell more internal success stories that can increase stakeholders 'confidence in GenAI implementation and show the practical benefits of AI-based solutions. These are vivid illustrations that show how the GenAI can help change many processes within an organization – starting with the organizational structure and ending with customer interactions, product design and ideas generation. To demonstrate the benefits of GenAI in the framework of the organization's activity, it is effective to emphasize specific cases where the program has produced a positive outcome.
- **Operational Efficiency:** An example of an internal success story could be that the company was able to use GenAI in eliminating repetitive functions such as data entry and report writing. Such tasks previously took a lot of time and with many mistakes made that delayed the overall process. This made productivity go high due to decreased time in completing tasks while staff directed efforts towards more important work. This also meant that the data compiled was accurate, thus helping in smarter decisions being

made. This story shows how GenAI creates new chances for making processes more efficient, cost-saving, and effective.

- **Enhanced Customer Engagement:** An example success story might state how a firm implemented GenAI-powered chatbots to reinvent the sale. These chatbots gave answer to customer queries promptly and accurately making it faster and satisfying for the customers. Therefore, the response times were reduced, the customer satisfaction markers improved and the chances of getting subsequent orders were enhanced. This example shows how might be useful GenAI in improving customer experience and hence strengthen the relationship and loyalty.
- **Innovations in Product Development:** A third success story could be an example of how organisation leveraged GenAI to push its product development agenda. GenAI was designed to identify markets and customer trends hence creating a set of recommendations with regards to a series of potential products that have not been realized or have not been popular among customers until now. This evidently resulted into achievement of the product launch and the attainment of market competition advantage. This example shows that using GenAI, organizations will be able to innovate and adapt to the needs of their consumers through the provision of products that meet the consumers' needs most closely. (Alwahedi, F., et.al,2024)

3. Clarify Limitations

Transparency About Limitations

Educational Materials: There is a need to create articles that point out to the public the limitations of what is being offered by generative AI (GenAI). GenAI contains numerous tools regarding automation, content creation, and even data analysis, however, it also harbours difficulties and threats. These limitations should always be captured in educational material about AI, to continuously remind people about the necessity of having human supervision and scepticism when working with AI products. Below is an elaboration on the key elements that these materials should include:

- **Potential for Errors:** It is necessary to develop articles relevant to the fact that the public needs to be informed about the limitations of what is provided by generative AI (GenAI). As we have seen, GenAI holds many tools when it comes to automation, content production and even data analysis; however, it also holds challenges and risks. These limitations should always be disclosed in educational material about AI to refresh

people's memory about the fact that the AI is constantly augmented by someone else, how it needs to be supervised by human being and the importance of scepticism when dealing with products based on AI algorithms. Below is an elaboration on the key elements that these materials should include: Here is an explanation about the content that should be contained onto these materials:

- **Limitations in Understanding Complex Contexts:** GenAI works great at learning data patterns and creating contents based on learned data, but it does not perform very well when it comes to applying contextual reasoning. This limitation can also be evident in other ways where the system makes response that are irrelevant to the context, basic or does not capture the shades of meaning. For instance, if a customer has a specific issue in customer service, they may be answered by a GenAI chatbot which will not be very helpful to the customer. Textbooks and other learning tools should always remind people that, again, GenAI has similar abilities to human brains regarding fake understanding and related feelings and nuances. There are circumstances when employees should be careful with AI and use professional assistance in different conditions that exist in the scenario.
- **Risks of Generating Biased or Inaccurate Content:** One of the main issues that can be attached to the GenAI approach is the possibility of arising content with the prejudice and mistakes in training materials. Potential sources of these biases should also be widely included in learning materials, along with the effects they have on the AI outcomes, the necessity in the use of varied datasets for training. There is a need to bring updates in the audit schedule, training processes, and training users to critically assess bias output in AI.
- **Ethical Considerations and Compliance:** The materials should also outline the ethical use of GenAI specially where GenAI is likely to be applied such as privacy, security, and decision-making aspects. For example, AI in creating content or recommendation can bring out an issue on data privacy and how users 'data are used, ethically. The educational materials should state the rules as to how GenAI should be used legally and ethically. This entails being upfront of the role that was played by AI in arriving at specific decision and to avoid creating content which will normally infringe on legal and / or ethical provisions.
- **Importance of Human Oversight:** It must be ensured that the use of GenAI must be under human supervision and this orientation must be highlighted in the educational materials. Being able to supply such computations autonomously has a great value,

although on the other hand human intervention and knowledge are the only adequate ways to analyse and interpret the results and make wise, accountable decisions. Users should cross check and authenticate the outputs of AI specifically where they are at a heightened risk. Materials should ensure GenAI effectiveness, efficiency, safety from error, bias and ethical issues to encourage right use besides ensuring that the strengths of the technology are fully exploited to give the maximum utility.

- **Scenario-Based Learning:** Scenario-based learning successfully showcases the GenAI weaknesses as the users get to experience different situations while using the technology. For instance, customer service chatbots may fail to understand questions which are asked to them, AI tools that recommend employees for a job may contain biases, and GenAI that is involved in legal issues or healthcare may have to be reviewed physically to avoid adverse outcomes. These visualises demonstrate why human intervention is required to prevent misuse of GenAI but also allows users to better understand its capacity for potent analysis.

4. Understanding Scope and Constraints

- **Workshops on Constraints:** It is quite essential to hold workshops where people can learn the challenges that come with the use of advanced Generative AI (GenAI) technologies. Such workshops would have been previously focused on what limitation GenAI has, where it terminates its functioning and the fact that these models are indeed powerful, but they are not perfect.
- **Dependency on Training Data:** This is one of the major drawbacks of using GenAI in that is bases its operations on the data that it is trained on. Indeed, in the workshops participants will be informed that biases or distortions in training data have an impact on biases or flaws in the AI results. For instance, recommendation systems that have been trained on historical data that has certain bias are likely going to have the same bias. Arguments will state that participants will view case studies and simulations in order to identify these bias and possible solutions such as using more diverse data.
- **Inability to Perform Tasks Beyond Programming:** GenAI technologies are pre-programmed and therefore are unable to work outside the set limits of their programming. This will be shown in the workshops where a text-generating model for instance will be unable to produce images. The audience will be taken through activities

that shall demonstrate the AI's weaknesses in handling new tasks. This is to point out that GenAI is an application that has its uses and is not a tool to replace human imagination and logical reasoning.

- **Challenges in Understanding Complex Contexts:** GenAI models often misinterpret complex or nuanced contexts, especially when these are underrepresented in training data. Workshops will show how AI may struggle with ambiguous inputs or subtle language differences, leading to incorrect responses. Participants will test these scenarios with interactive examples, highlighting the AI's limitations in understanding context and the need for human oversight to ensure accurate and relevant content.
- **Risk of Generating Inaccurate or Biased Content:** Another major drawback that would be addressed in the workshops includes the possibility of GenAI to produce incorrect or prejudiced information. The audience would get to know that the utilization of AI generated outputs need not be fully trusted. For instance, a GenAI tool applying to legal document writing may come up with a statement that is legally incorrect or does not meet the needs of a particular case. In stimulated cases, participants would learn about examples of failed AI content and the consequences of AI-generated content and would identify methods of how AI content may be reviewed by humans. The workshop would also pay special attention to the fact that the training of AI models should be continuous and updated according to the new data and new standards.
- **Limitations in Predictive Accuracy:** GenAI is an effective tool to make predictions, but these predictions are not always accurate, and their accuracy relies on the quality and the extent of the data used by the system. The workshop would discuss how the GenAI models may produce the accurate and correct output from the incomplete or the stale data which may lead to wrong decisions. It would involve activities that would involve participants to predict the outcome of certain situations based on the AI and then compare it with the actual outcome and how over reliance on such predictions could lead to hazards.
- **Ethical Considerations and Responsible Use:** Last but not the least, the workshops would address issues such as the rights and wrongs in the application of GenAI technologies, the risks of misuse of the technologies as well as the importance of proper use of the technologies. It will be proposed to discuss the cases where the usage of AI may be problematic, like surveillance, decision making, or content generation, and to consider the approaches to control how AI is used in a proper manner that meets the company's values.

Therefore, the participants would be able to acquire an understanding of the scope of limitations of the GenAI technologies in these workshops. This is important to avoid incorporation of GenAI in the work process in a way that would make it improve efficiency and decision making yet at the same time compromise on the accuracy, fairness, and ethical implications.

5. Interactive Q&A

Enabling live questions and answers that can be asked by the stakeholders regarding the constraints of Generative AI (GenAI) is one of the most important ways of building a profound and applied understanding of the technology. These sessions provide an opportunity for the stakeholders to interact with the AI professionals and get their doubts cleared and address some of their concerns about the use of GenAI in their organisations.

- **Structured Q&A Framework:** Such questions and answers will be arranged in a way that the sessions will be informative, and they will allow the participants to ask questions freely. Participants would be expected to come to the session with some questions especially those that are pertinent to their functions and responsibilities at the start of the session. This format helps in maintaining a proper direction and flow of the session and addresses the most important concerns. The facilitator would then take charge of the discussion and make sure that all the questions are answered in a systematic manner starting with the general questions and then to the specific ones.
- **Real-World Examples and Case Studies:** It is important to note that the AI experts would have answered the questions and queries during the Q&A sessions by giving real-life examples and case studies. For instance, when a stakeholder inquires whether there is a possibility of AI to bring a bias in decision-making, the expert may then explain a situation where AI recommendation turned out to be biased based on the data that was fed into the system. They will then describe how those biases were discovered and addressed and what recommendations the organization can make to ensure that such occurrences are not repeated. Not only does this approach answer the question but it also provides the stakeholders with recommendations that they can implement in their own context.
- **Deep Dives into Specific Limitations:** Some of the questions that the stakeholders may have include; limitations of GenAI such as the fact that it is data dependent, it does not fully comprehend complex contexts and issues with integration of AI in the current

workflow. The Q&A sessions would enable discussions on these issues to a certain extent while experts could explain the concepts with technical understanding and real-world examples. For instance, if a stakeholder is worried on how AI could mislead regarding the language used in customer's queries, the expert could reveal that natural language processing techniques have their limitations in this aspect and recommend how the performance of AI can be improved through the addition of human supervision or more training data.

- **Tailored Responses for Different Stakeholders:** The interactive sessions will be developed in a way that will benefit different participants from the technical staff, who requires practical information, to the non-technical business personnel. For instance, technical people may be concerned with the details of algorithm adjustments or how best to fine-tune AI tools for increased efficiency while businesspeople may wish to know how the limitations of AI affect business results. It is beneficial to present the information in the Q&A format as this way the speakers can address each participant's concerns regarding the AI limitations in their workplace and how to overcome them.
- **Encouraging a Collaborative Learning Environment:** These Q&A sessions would also create a platform for group learning since all the stakeholders will be free to share their experiences and ideas. GenAI tools are very effective, and people would be asked to bring out issues they have encountered or may encounter while using GenAI tools this would make the discussion more detailed. This group discussion in which the expert plays a role of a moderator assists in identifying the general concerns and, therefore, the effective practices may be easily implemented across the organization. For instance, a business unit that has already embraced AI may offer tips that it has gathered in the process, which will be useful for other units that are only starting to consider integration of AI.
- **Addressing Ethical and Practical Concerns:** Besides, there are factors like ethical or practical considerations which may hinder the adoption of GenAI depending on the specific application such as privacy concerns, unfairness, or job loss.
- Other concerns that can also be classified as non-technical barriers include; Ethical considerations such as privacy or fairness; Practical considerations such as the fear of the loss of jobs among others. These concerns are best taken to the Q&A sessions where all the participants can express them freely. They can provide objective opinions and insights and present both the advantages and disadvantages of AI as well as measures that may be taken to prevent the negative applications of AI. This can be done by talking

about the principles that have been put in place to regulate the use of AI or the principles that need to be followed to make the AI systems to be more accountable in their operations.

- **Continuous Engagement and Follow-Up:** To make sure that the effects of the Q&A sessions are not limited to the first discussion there would be methods of further participation and feedback. Further questions could be asked by the stakeholders after the session, and these would be answered in the following meetings or in writing. This ongoing conversation serves to further support learning and offers stakeholders a way to effect continuous improvement in their comprehension of AI's capabilities and constraints as they go about their work with these technologies.

Thus, through the implementation of the interactive Q&A sessions, it will be easier to address the current and future issues of GenAI and ensure that all the stakeholders are in a position to understand the impacts of GenAI and how to make the best use of it without exposing them to the negative consequences of using it.

5. Implement Best Practices

Ethical Use and Governance

- **Policy Development:** Thus, it is necessary to focus on the creation and promotion of the policies that define ethical usage of Generative AI (GenAI). It is therefore important that these policies should cover on aspects like transparency, accountability and fairness in a bid to ensure trust and avoid possible dangers. Transparency is the act of offering information on the functioning of the AI systems, including the procured data, the algorithms and the decision-making process. This helps users to know the fundamental of AI outputs and hence they can trust the technology. Accountability means that there should be a clear description of who will oversee and manage the output of the AI and how the reporting process will occur in the case of identified problems or bias. In this way, organizations can know who oversees managing how the AI operates and how to handle malfunctions so that AI is well implemented in the right manner. This means that to achieve fairness in AI it is necessary to take measures against and eliminate biases in the AI decision making. The following measures should be set as policies; Conduct periodic assessments of the algorithms to identify areas of bias and make the necessary adjustments, use different data sets and; Ensure that algorithm development includes people from all backgrounds. Thus, defining these guidelines and making sure

that the organizations follow them properly through training and regular checks, organizations can promote the ethical use of AI and therefore the positive impact of AI.

- **Ethics Training:** It is therefore important to provide ethics training to the employees to make them understand the ethical issues that come with the use of Generative AI (GenAI) and thus make it possible to deploy it responsibly. Such sessions should start with an introduction to ethical issues of GenAI such as the issue of misinformation, bias, and privacy. Moral concerns with the use of AI should be addressed by first training the staff on how to avoid creating negative or fake texts, second, creating guidelines that would help in avoiding the creation of such content and third, review of AI outputs for any form of undesirable content. Other topics that should be included should also include proper handling of data, privacy issues, and legal and ethical policies on the use of AI. Real and hypothetical ethical dilemmas can be presented which can help the participants to learn about some of the ethical issues which they are likely to encounter and how such issues can be addressed. This is because employees are given refresher courses frequently and are provided with constant support to make them understand the new ethical concerns regarding the use of Artificial Intelligence. Thus, this approach to ethics training allows eliminating the risks of creating improper content and supports the ethical utilization of GenAI within the company.

6. Data Security and Privacy

- **Security Protocols:** It is therefore imperative to train the stakeholders in the matters concerning data security and privacy to avoid leakage of information and to prevent breach of data protection policies. Training should therefore emphasize on giving clear policies and measures on the management of data and protection of privacy.
- **Handling Sensitive Information:** Some of the possible areas that stakeholders should be trained on include; how to manage; personal information, financial information and health information. It should be more on how to ensure that this data is stored, retrieved, and transmitted in a right manner with an aim of enhancing security. Some of them include, employing the highest levels of encryption, safe ways of storing data, and restricting access to the data to only authorized persons.
- **Securing Data Transfers:** It is therefore important for stakeholders to understand several measures that can be used in protecting data in transit. Training should include guidelines on the use of safe communication tools for instance encrypted emails and safe transfer of files/solutions (for instance SFTP or HTTPS). All the stakeholders

should also be informed on the need to confirm the identity of the intended recipient and the need to use strong authentication technique in the transmission of the data. In the same way, training should involve the use of data loss prevention (DLP) tools and ways on how to manage data transfer.

- **Complying with Data Protection Regulations:** It is important that organizations must know and follow the rules regarding data protection so that they will not violate any legal policies and to build the trust. Training should give a general idea of the legal requirements, for instance, GDPR and CCPA. The following are some of the fundamental principles that the stakeholders should know include; consent that is getting consent to collect data, data subject rights for example right to access, right to rectification, right to erasure, and data breach notification. Training should also consist of the guidelines for compliance, for example, keeping a record of processing activities and conducting checks.
- **Implementing Data Privacy Policies:** It is therefore important that any training that is to be conducted should incorporate the following: On how to formulate and implement policies on data privacy within the organization. The public should be informed on the need to have well defined policies that address issues to do with data usage, protection mechanisms and privacy. Such training should include aspects such as; how to design and draft such policies, how to assess and revise such policies from time to time in line with the changing legal requirements and practices. It should also be ensured that the employees are well informed on how to convey such policies within their respective departments.
- **Handling Data Breaches and Incidents:** Training should equip the stakeholders to deal with data breaches and security incidents appropriately. This includes apprising them on the measures that should be taken in case of a breach including communicating the breach to the affected persons, containing the breach and communicating the breach to the appropriate authorities. Some of the other measures that the stakeholders should also be put through include; how to carry out an incident handling process, how to analyse the causes of the incidents and how to work on the corrective measures to avoid occurrence of similar incidents in the future.
- **6. Continuous Improvement and Updates:** The issue of data security and privacy cannot be discussed without mentioning that updates should be made periodically hence; stakeholders should be encouraged to seek information on the changes and the recommended precautions. There is also a need to incorporate on-going training of new

threats, changes in regulation, and new technologies that are available in the market. Ongoing education and resources, it is possible to ensure that all the stakeholders are well informed and ready to face new and changing security and privacy threats.

- **Privacy Workshops:** Privacy workshops are crucial tools in the protection of privacy when it comes to the implementation of Generative AI, also known as GenAI. They include methods of data anonymization, model protection, and privacy techniques including differential privacy. The participants will also be able to understand how to adhere to the laws such as the GDPR as well as the CCPA by employing tools such as the privacy impact assessments. The workshops include such activities as simulations, group discussions, and case studies to work on privacy-related issues and to get support and updates on the best practices and rules.

7. Bias Mitigation

- **Bias Awareness:** It is therefore important to provide the Generative AI (GenAI) with bias awareness training for ethical use of the technologies. It starts with the discussion of biases which are algorithmic, data, and societal, and their source in the training data and model creation. The training also offers materials namely checklists, diagnostic tools, and cases to help in identifying biases. The employees are taught on ways of assessing the impact of bias on the AI results including aspects such as the reliability and equity of the results as well as they are taught on how to monitor AI systems regularly and update them to counter new forms of bias. This approach helps in making the AI systems non-biased and more accurate and in compliance with the set ethical principles.
- **Diverse Data:** Training for Generative AI commonly known as GenAI has called for the use of data that is inclusive to everyone to ensure that the system does not bias and enhance the AI performance. Key areas include:

  1.Importance of Diverse Data: Select data that can represent diverse demographic, cultural, and contextual factors to avoid stereotype and meet the needs of all the users.

  2. Bias Evaluation: Identify the cases of lacking representation in the datasets and apply data balancing or augmentation techniques to rectify it.

  3.Bias Mitigation Strategies: Do things like use fairness-aware algorithms, checking for bias, and techniques such as adversarial debiasing. It is, therefore, necessary to keep on monitoring the situation and making necessary changes.

This helps to make the AI models to be fair, accurate and incorporate for the better performance of the models.

8. Promote Continuous Learning

Ongoing Training

- **Regular Updates:** It is therefore important to refresh the staff's knowledge on the developments in Generative AI (GenAI). Routine meetings should include the introduction of new features, development and trends in the field, and should be conducted in a more engaging manner such as through webinars and workshops. These updates assist the staff to integrate the recent changes that have occurred in the given field into their practice. It also means that training should also cover for instance the frequent audits of the existing AI systems as to check their relevancy and efficiency. Through this, the organizations can prepare the employees for changes in technology that may arise, and thus adequately manage the AI systems.
- **Webinars and Seminars:** Consequently, frequent webinars and seminars featuring Generative AI professionals should be conducted to make sure that employees are well informed on the latest trends and techniques. Webinars are great for those wanting to watch the latest developments in GenAI algorithms at their convenience while seminars are more formal, long and involve presentations followed by questions and answers. All the formats must incorporate real life examples and examples so that the employees can implement what they have learnt. These sessions help staff to be up to date with the latest advancement in AI, contribute to the ongoing improvement, and improve the organization's strategic application of GenAI.

9. Feedback and Adaptation:

- **Feedback Mechanisms:** To enhance GenAI tools and practices, many sound feedback systems must be put in place. The following ways should be used to capture staff's feedback: Surveys, suggestion forms and feedback meetings. Surveys provide quantitative data on the usability and the relevance of the training to the organization while the suggestion forms provide qualitative data. Such sessions allow participants to

share opinions and ideas and provide real-life examples. Feedback analysis allows to see trends and possible problems that may occur during training as well as to update training materials and best practices. This recursive approach ensures that GenAI tools are relevant and are able to solve new challenges as they arise and at the same time improve on their performance.

- **Continuous Improvement:** Continuous improvement process of GenAI tools entails the collecting and analysing of feedback from the staff to enhance the policies, training materials and practices of the tools. This is because feedback should be reviewed at regular intervals to be able to make certain observations and improvements. New policies should be developed if there are issues that need to be addressed and if policies are to be in line with current practices in the industry, then they need to be revised. It also takes time to change practices and workflows, and such changes should be made based on feedback to enhance the efficacy and efficiency. Checks on the progress of the process are conducted on a regular basis with a view of ensuring that the process is still relevant and up to date with new trends and ideas, with the staff playing an important role.

10. Foster a Culture of Awareness

Awareness Campaigns

- **Internal Communications:** To disseminate information on Generative AI (GenAI) within an organization, internal communications campaigns must be conducted. These campaigns are very important in the sense that it will help to create awareness to the staff in the organization on the features, the weaknesses and the right ways of applying GenAI, this will ensure that everyone is in grasp with the organizations AI strategy.

- **Newsletters**: It is quite useful to send regular newsletters that could help in promoting information about GenAI. These newsletters should be used to inform the subscribers about the new advancements, new features, and new guidelines. These may be articles from the GenAI in-house writers or guest writers, real life experiences of organizations that have adopted GenAI and how they have done so and getting started guides on how to use the GenAI tools. Thus, by providing such information in a regular basis, newsletters serve to remind GenAI's staff and keep them informed of latest developments.

- **Emails:** The more specific and timely information about GenAI can be shared with the customers, targeted email campaigns can be used for this purpose. For instance, messages may be used to introduce new tools, emphasize on new changes or simply remind the members about the next training session. Personalized emails can also be sent to departments within an organization for instance technical staff or project managers with content that is suited to them.
- **Company-Wide Meetings:** Team meetings are very effective when it comes to holding meetings face to face and they can be used in creating awareness of GenAI. Some of the topics that may be discussed during the said meetings may be the current development in GenAI, accomplishments that have been made, and visions for the future. Presentations could be by internal or external experts while question and answer sessions provide an opportunity for the staff to ask questions and make their contributions. It can also be a venue for providing information on the consequences of using GenAI tools and acquire feedback from the employees.
- **Interactive Platforms:** Some of the ways of sustaining the discussion include working through internal forums or chat groups. These are means by which the staff can share their experiences, raise their queries and even give their opinions and feedback on the spot. They also assist in sharing other information that may include guidelines and video tutorials that may help the staff to use GenAI tools. **Campaigns and Initiatives:** Others may include GenAI Awareness Week, or AI Innovation Days to ensure that the employees get to know and appreciate GenAI and its importance in the organization. Such campaigns can comprise of several strategies such as workshops, demonstrations and contests/ competitions which can be useful in portraying the strengths of GenAI. It also assists in clarifying any misconceptions as well as promote the right practices among the students (Bada et al., 2019).

  Therefore, through newsletters, emails, company meetings, and social media application, the teams can be informed of GenAI and the areas it can handle, the areas it cannot handle, and how it should be used. These communication strategies help in developing the shared understanding of what GenAI is and how it should be used as well as how AI can be incorporated into business processes of the organization.

- **Promotional Materials:** Using posters, infographics, and other outputs is an effective method of reminding the employees about the concept of Generative AI and increasing the awareness of the term within an organization. These visual aids act as references

that contain important messages and guidelines to follow to ensure that such information is well imbibed by the workers.

- **Posters:** Create effective posters which will focus on the key features of GenAI, strengths, and weaknesses, as well as guidelines for application. These posters may contain short and simple messages with the use of illustrations, graphs or bullet points. Posting these posters in key areas like break rooms, corridors, and areas where most of the workers are likely to be will help staff be constantly reminded of GenAI hence enhancing their learning.
- **Infographics:** Create posters that can explain the various information about GenAI in simple and attractive diagrams and graphs. The various concepts about GenAI can be depicted in infographics, examples of where and how GenAI can be applied and important considerations to make when using it. They can also illustrate things one has to refrain from doing and things to do in order to be effective. Infographics should have readability that does not require much time to comprehend so that one can easily refer back to them.
- **Promotional Materials:** Utilize various promotional materials, such as flyers, bookmarks, and digital screens, to share information about GenAI. Flyers and bookmarks can be distributed during meetings or training sessions, while digital screens in common areas can display rotating slides or videos about GenAI. These materials can include QR codes linking to additional resources or training modules, providing staff with easy access to more in-depth information.
- **Interactive Displays:** It may also be useful to establish so called touchpoints or info-points in the form of interactive screens or kiosks to be placed in such locations where representatives of the GenAI have a possibility to communicate with the audience. Such screens can also have touch screens or tablets through which the employees can go through some tutorials, watch brief instructional videos, or answer quizzes concerning GenAI. The use of the interactive displays is effective since they make the learning process to be more enjoyable while at the same time they can accommodate the different learning modalities.
- **Display Strategy:** To make sure that these materials will be effective, it is recommended to put them in areas where employees may be found or areas that they tend to visit often. This make the information to be seen by everyone so that they can

be able to access it. Make changes and revisions as and when new information becomes available so that the materials remain current and useful.

Through the generation of posters, infographics, and other promotional materials, organizations can help keep knowledge about GenAI fresh and easily seen by the staff and thus improve understanding of it as well as its strengths, weaknesses, and how it should be used. This approach ensures that the flow is sustained, and gen AI knowledge is well incorporated into the organization's culture, thus improving the utilization and application of AI technologies.

11. Collaboration and Dialogue

- **Cross-Functional Teams:** It is imperative to engage the IT, marketing, compliance and other departments to tackle some of the emerging challenges with the GenAI. By having inter-departmental meetings, creating task force and building knowledge base, organizations can encourage exchange of ideas and group discussions. Combined training sessions and workshops are effective in ensuring that the teams learn each other's roles and to appreciate the different points of views. Feedback systems and the positive culture of tolerance for people also help coordinate the processes even better. This approach helps in covering all the aspects of GenAI and thus comes up with better and well-coordinated implementation strategies.
- **Knowledge Sharing:** Developing a platform to discuss situations and concerns of the workforce regarding Generative AI (GenAI) increases collaboration and makes the staff aware of the concepts of AI. These forums offer a chance for the employees to share their experiences, resolve issues that affect them and even learn from each other on the best practice to follow. In this manner, organizations can tap into the different opinions and ideas and share them towards the achievement of better decision-making when it comes to GenAI implementation and management. This way of working makes the overall improvement of AI technologies and increases the possibilities of organizational learning. (Dwivedi et al.,2024)

It is imperative that there is collaboration between cybersecurity professionals, data scientists and business personnel in order to optimize the use of Generative AI (GenAI) while at the same time maintaining adequate security and business relevance. In this way organizations can coordinate the efforts of these groups and utilise the collective wisdom of GenAI.

It will also help to have Cybersecurity Experts to help in giving recommendations on how best to secure AI systems and data. This is because the professionals involved in the creation of AI tools understand the potential threats that are involved such as hacking or theft of data.

Data Scientists support the development of GenAI models by applying their expertise in AI algorithms and big data to refine the GenAI models' performance. Their work may include adjusting AI models and making sure that it is working optimally as well as following the recommended guidelines in handling data.

Business Stakeholders give insights for how GenAI can be implemented towards the achievement of the company's goals as well as meeting the consumers' demands. Their input enables the development of AI solutions that are relevant to the business needs, Customer Experience and Innovation.

This way the organizations can guarantee that the cybersecurity concerns, data science best practices and business strategies are well aligned. This way, GenAI tools not only get better security and performance, but the organization can also be assured that the GenAI tools add value to the organization and align with the goals of the organization.

2.4.8 AI Training Program

Step-1

Foundational Training

- – Introduction to Generative AI concepts
  – Key AI tools and their applications
  – Audience: All employees

Step-2

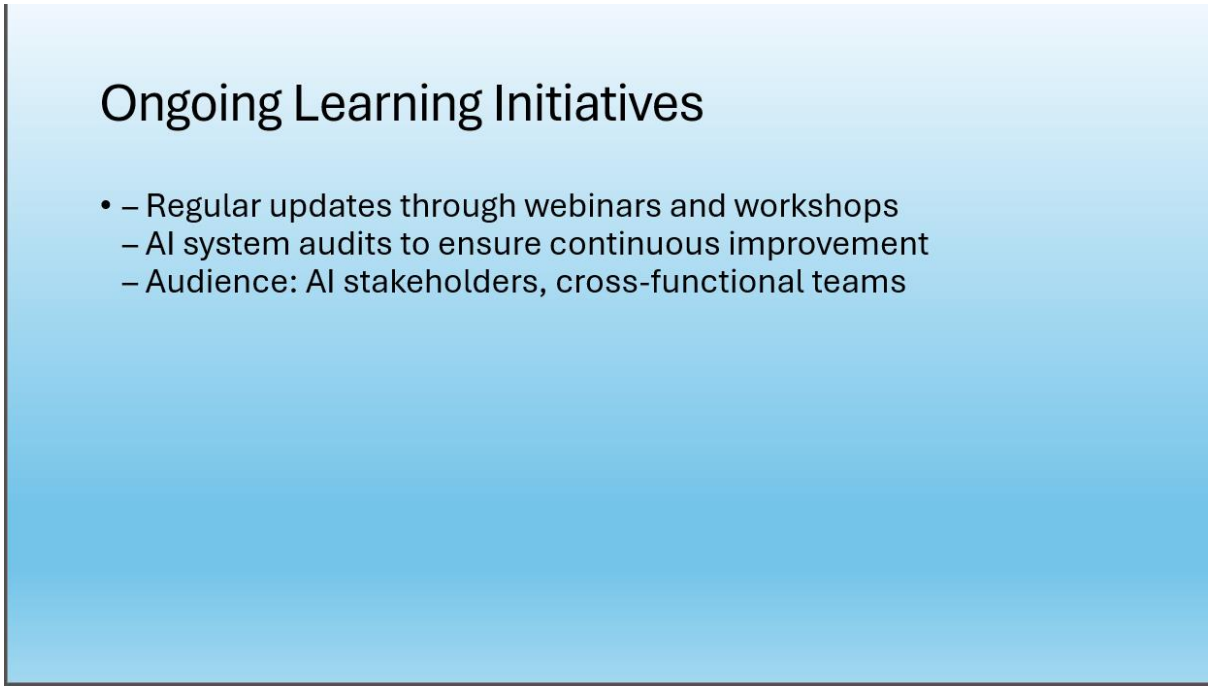
Technical Mastery Sessions

- – Deep dive into AI development, algorithms, and coding (Python, AWS)
  – Practical workshops on AI deployment
  – Audience: IT, developers, data scientists

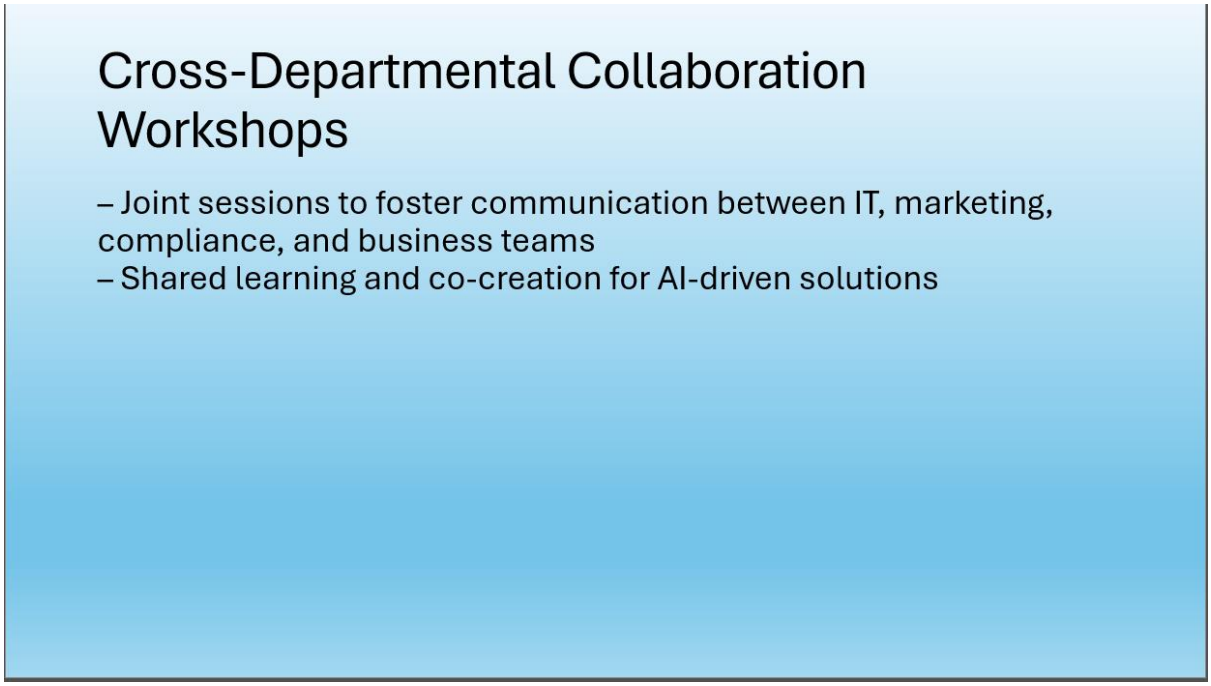
Step-3


Ethical AI & Compliance Training
• – Understanding AI ethics, data protection, and legal frameworks
– AI governance and responsible AI use
– Audience: Legal, compliance, and management teams

Step-4


Ongoing Learning Initiatives
• – Regular updates through webinars and workshops
– AI system audits to ensure continuous improvement
– Audience: AI stakeholders, cross-functional teams

Step-5



Step-6

The AI Training Programs are designed to ensure that everyone in the organization (automobile service industry) is well-prepared to work with generative AI. By providing foundational training for all employees and offering more technical sessions for IT and development teams, we ensure that expertise is built at every level.
The Ethical AI & Compliance Training ensures we stay aligned with legal and ethical standards, fostering responsible AI use. Continuous learning, through regular updates and system audits, keeps our teams up-to-date with the latest advancements.
Lastly, the collaborative workshops across departments will encourage innovation and shared learning, helping us leverage AI to drive organizational growth. Together, these training programs will create a robust, AI-ready workforce.

2.5 Website design and development

Regarding the technical progress of the project till now, this project has developed and designed a website for Automotive industry specific to WaterTunnel car wash company. For developing websites, deployed Wordpress certified by Bitnami by using AWS web server involving several steps. The main reasons for choosing AWS cloud service for deploying WordPress are security and cost-effectiveness. When it completed the process of launching instances in EC2 it received a public IPv4 address and login username and password for the WordPress website.

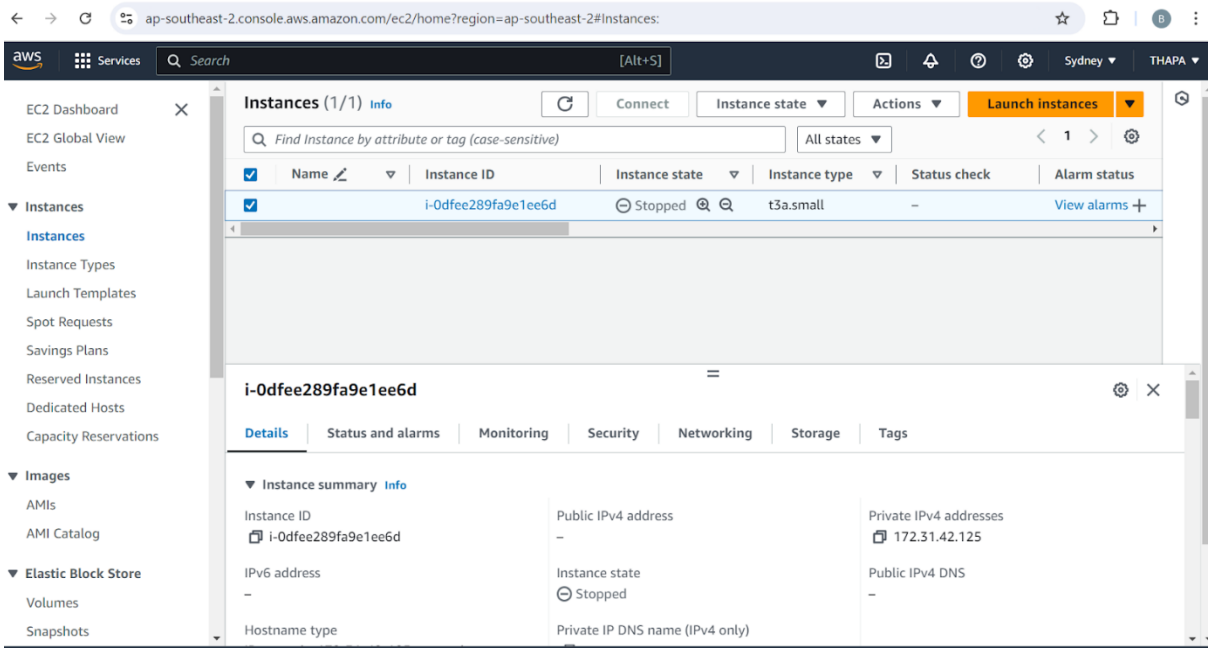


Figure-11: Launching Bitnami WordPress in AWS.

When creating websites, collected different resources from online sources and some of them are self-designed. The website contains everything that needs to be a perfect website. The website is all about company's information and car wash features. The website includes service details, Wash Menu, contact details etc. Here are some glimpses on the websites.

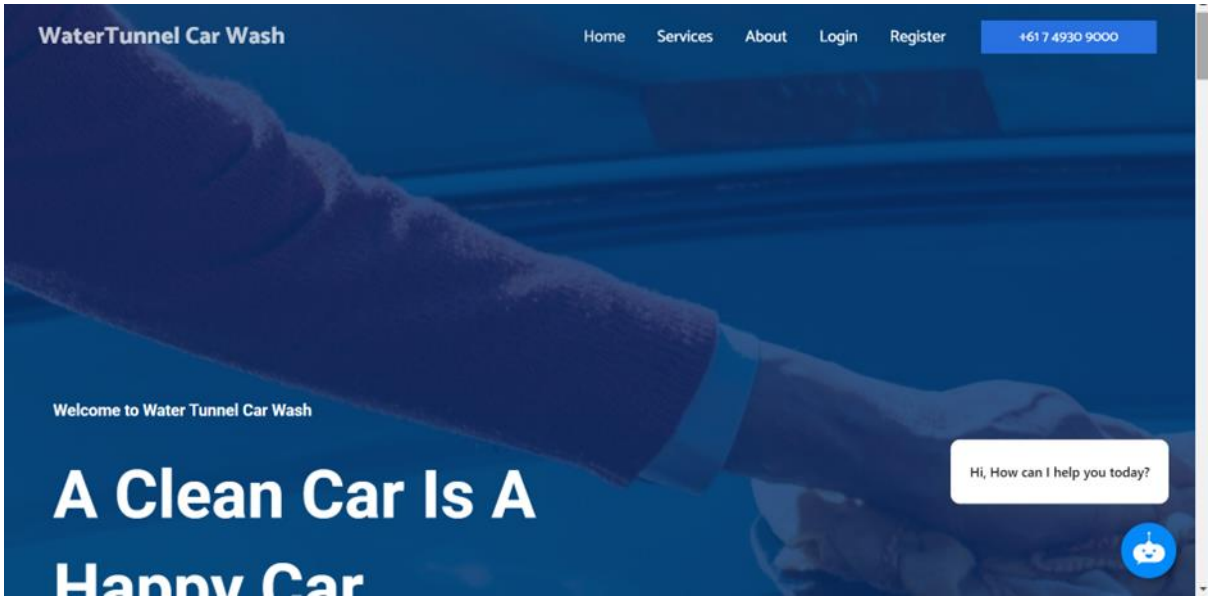


Figure-12: Company Website Dashboard

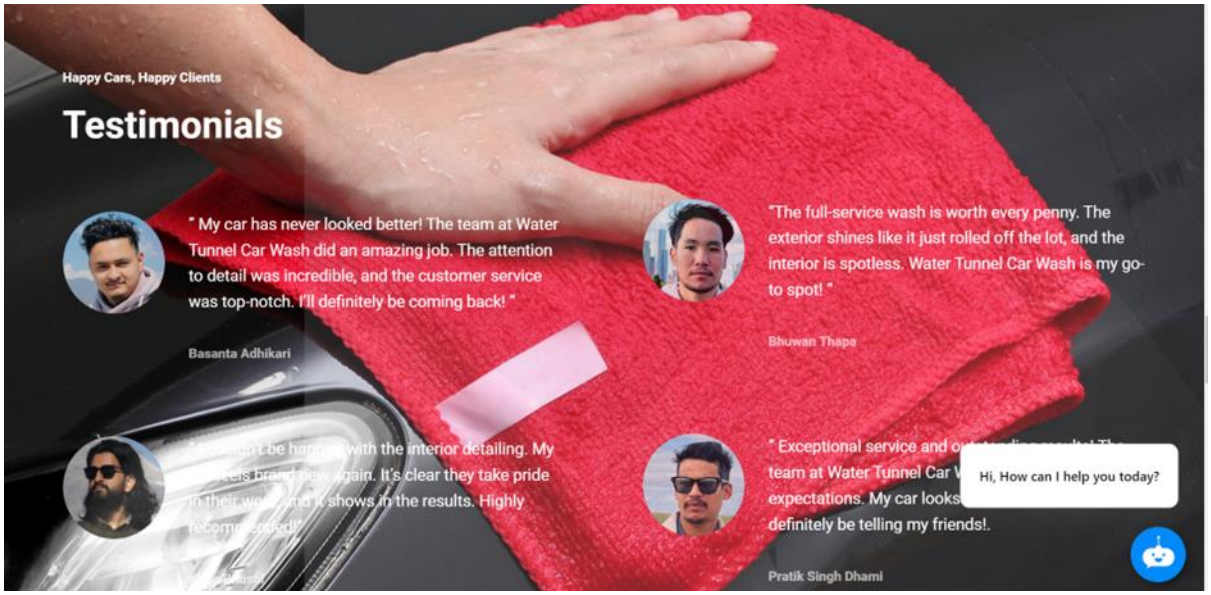


Figure-13: Company Website

2.5.1 Deployment process of WordPress website

The developed website is deployed into EC2 instance under AWS server for scalability, reliability, security and cost-effectiveness etc.

Deployment process of website in EC2 instance:

1. Login to AWS server.
2. Select “EC2” instances from all services.
3. Select “lunch instance” under EC2.
4. Select “browse more AMIs” by default.
5. Select “AWS Marketplace AMIs” under AMI.
6. Select desired WordPress website.
7. Name website “WordPress or any other name” under key pair.
8. Select “lunch instances” at the bottom and wait for couple of minutes
9. Again, go back to instances under EC2, you will see running instances
10. Select “running instances” you will see details of the website such as public IP and private IP.
11. Copy “public IP” and open in new tab you will see website format and design it.
12. For login username and password for WordPress, select running “instance ID”.
13. Select “action” and again select “monitor and troubleshoot”.
14. Seach for user Id and password to login in WordPress site.

2.6 AI Tool to Detect Vulnerabilities in Cybersecurity

Gen AI plays an important role in providing security layers for the website or any other application. As we already discussed about the problems, solutions, benefits and limitations of GenAI while detecting the vulnerabilities in cybersecurity. Here, our aim is to generate such AI tool which detects Vulnerabilities on the website that we already designed.

The AI will be based on the following algorithms:

1. Data collection:
    - Gather website content including HTML, CSS, JavaScript and backend code for analysis.
    - Extract security related metadata such as headers, cookies, and response codes.
2. Input processing:
    - Tokenize and purse the website code into analysable units.
    - Normalise data to eliminate any inconsistencies.
3. Vulnerability detection:
    - Pattern matching: Identify known vulnerability pattern using predefined rule sets (e.g SQL injection, XSS, CSRF).
    - Anomaly detection: Use AI/ML models to detect unusual patterns or deviations in the website's structure and that may indicate new vulnerabilities.
4. Security Compliance Check:
    - Compare website configurations and code against industry security standards (e.g., OWASP Top 10) to ensure compliance.
    - Generate a compliance report highlighting areas of concern.
5. Recommendation Generation:
    - Based on detected vulnerabilities, generate actionable security recommendations.
    - Prioritize vulnerabilities by severity and impact.
6. Reporting:
    - Generate a detailed security report with findings, risk assessments, and recommended actions.
    - Provide options for automatic or manual fixes based on severity.

7. Deployment:
    - Integrate the Al tool into a continuous integration/continuous deployment (CI/CD) pipeline.
    - Deploy the Al tool on AWS to enable scalability and access control.
8. Feedback Loop:
    - Continuously update the Al models with new vulnerability data and improve detection algorithms based on user feedback.

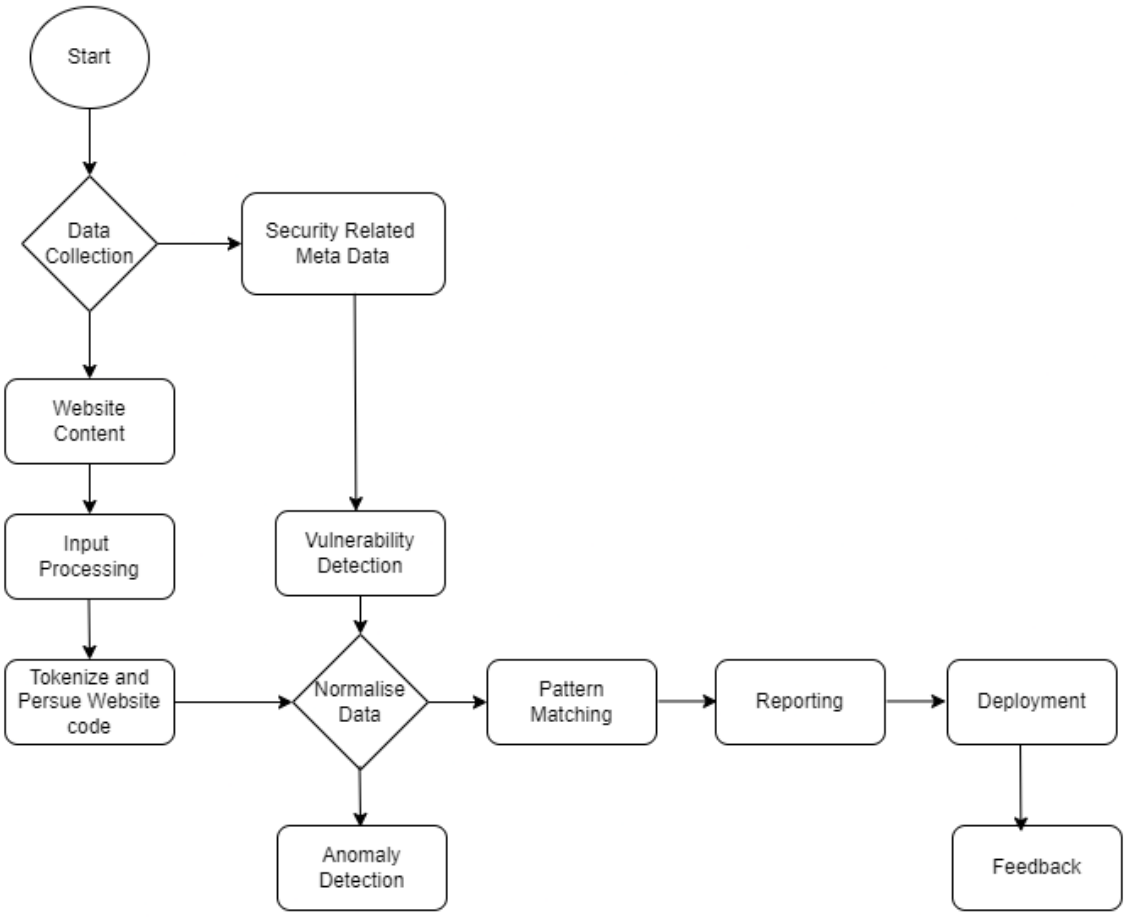


Figure-14: Flowchart for GenAI Tool

2.6.1 Software Requirements Specifications (SRS)

Here the topic defines the functional and non-functional requirements for the web application vulnerability scanner. The system is developed to scan website for common vulnerabilities like SQL injection, cross-site scripting (XSS), SSRF, RCE and many more. The scanner will help developer and security leaders identify and assess vulnerabilities in their website.

The web application vulnerability will allow user to enter a URL and scan various security vulnerabilities of the website. The system checks foe vulnerabilities across common endpoints, Analyse server responses and classify the results. The vulnerabilities logged and reported in the comprehensive format. The scanner also performs brute force login attempts and anomaly detection in server responses.

SQL Injection: A code injection technique that allow unauthorized access to a database, enable attackers to view modify and delete data.
XSS (Cross Site Scripting): When web application accepts user input and includes in webpages without proper validation or sanitization the malicious script is permanently stored in webpages.
Server Site Request Forgery (SSRF): When a user clicks on vulnerable URL, that allows attackers to access the information of the users.
Cross Site Request Forgery (CSRF): An attack where an attackers trapped users to perform actions on web application by sending them links.
Brute force Attack: an attack that used to gain unauthorised access to the system by systematically trying every possible combination of password.

Functional Requirements

1. The system must accept URL as input from user.
2. The system will check common paths such as s (e.g., /wp-admin, /wp-login.php, /phpinfo.php) for vulnerabilities.
3. The system will classify the type of vulnerability using a pre-trained machine learning model such as (e.g., SQL Injection, XSS, CSRF, etc.).
4. The system will attempt brute force login with common username and password combinations.
5. The system will detect anomalies in the server responses based on a trained anomaly detection model.
6. The system will log vulnerabilities and errors to an HTML file.

Interface requirements

1. User interface: valid URL for scanning and HTML pages showing vulnerabilities results, classification and report.
2. No hardware requirements.
3. Software requirements: python, flask that serves as a web framework, TensorFlow used for machine learning classification, Hugging Face Transformers for pre-trained model integration.

2.6.2 Deployment of AI tool onto AWS server

The developed AI tool is deployed into EC2 instance under AWS server for scalability, reliability, security and cost-effectiveness etc.

Deployment process of AI tool

1. Prepare AI model
- Choose framework e.g TenssorFlow

2. Create deployment package

- Setup virtual environment

Type: python -m venv myenv

source myenv/bin/activate

- Install necessary libraries

Type: pip install tensorflow

pip install numpy

- Package Your Code: Create folder and copy AI model and code

Type: mkdir my-deployment-package

cp -r mymodel.h5 my-deployment-package/

cp lambda_function.py my-deployment-package/

cd my-deployment-package

zip -r ../my-deployment-package.zip .

cd ..

3. Login to AWS management console

4. Navigate to lambda

- Click on "Create function."

- Choose "Author from scratch."

- Fill in the function name, select the runtime (e.g., Python 3.x), and set execution permissions (IAM role).

5. Upload Your Deployment Package

- In the "Function code" section, select "Upload a .zip file" and upload my-deployment-package.zip.

6. Set necessary environment variables in lambda configuration.

7. Increase memory and timeout if necessary.

8. Go to “Test Tab” and create a new test event.

9. Run the test and check the output.

Steps for Running AI tool

1. Login to AWS server.
2. Access the EC2 instance.
3. Navigate to AI Scanner:
- Select the EC2 instance that host AI Scanner.
4. Start the instance:
- If the instance is in a stopped state, go to the instance state and click start.

5. Connect the EC2 instance:
- Once the instance is running, select the instance and click connect.
6. Establish a connection:
- Choose the method to connect to your instance (EC2 instance connect or other) and click connect at the bottom to initiate the session.
7. Navigate to Web Scanner Directory:
- In the terminal type the following command to navigate to the Web Scanner directory: cd Web_Scanner/
8. To see deployed python code type: cat app.py
9. Activate the Virtual Environment by typing: source venv/bin/activate
10. Start the Web Server:
- Start the AI Scanner application using Gunicorn with following command: gunicorn -w 4 -b 0.0.0.0:5000 -timeout 1200 app:app

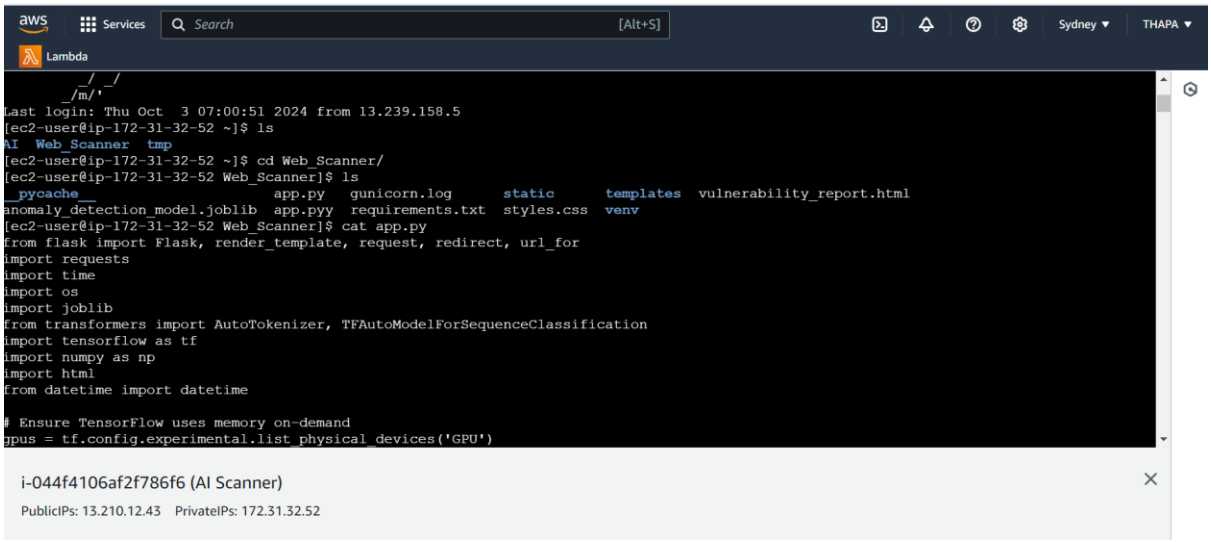


Figure-15: Running AI Scanner in Virtual Environment on EC2.

2.6.2 AI to Identify Cybersecurity Vulnerabilities

This prototype completed the basic task of developing AI which is able to detect cybersecurity vulnerabilities, and checked using URL of the website and got positive result. The AI is basically generated in VS code using python programming language, also used the basic models available on python library and used AI ML model such as isolation forest and tensor flow for anomaly detection and also used pre trained AI model Microsoft/codebert-base for making task easy. Basically, the generated AI can detect vulnerabilities on the different areas of the websites through SQL injection, Cross site scripting, brute-force attack and AI tool also tried Cross-site Request Forgery (CSRF) and Server-site Request Forgery (SSRF). The result has been shown below.

Vulnerability Scan Report for http://52.63.41.222/

Path	Description	Severity Level
wp-admin/install.php	[VULNERABLE] http://52.63.41.222/wp-admin/install.php - WordPress installation script is accessible. (Type: XSS, Confidence: 0.55)	
wp-admin/install.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-admin/install.php with payload: http://localhost [ANOMALOUS RESPONSE DETECTED]	
wp-admin/install.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-admin/install.php with payload: http://127.0.0.1	
wp-admin/install.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-admin/install.php with payload: http://169.254.169.254/latest/meta-data/ [ANOMALOUS RESPONSE DETECTED]	
wp-login.php	[VULNERABLE] http://52.63.41.222/wp-login.php - WordPress login page is exposed. (Type: XSS, Confidence: 0.56)	
wp-login.php	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-login.php with payload: <script>alert('XSS')</script>	

Figure-16: Vulnerability Scan Report

wp-login.php	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-login.php with payload:
wp-login.php	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-login.php with payload: <div onmouseover='alert("XSS")'>Hover</div>
wp-login.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-login.php with payload: http://localhost
wp-login.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-login.php with payload: http://127.0.0.1
wp-login.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-login.php with payload: http://169.254.169.254/latest/meta-data/
wp-login.php	[VULNERABLE] RCE possible at: http://52.63.41.222/wp-login.php with payload: phpinfo()
wp-login.php	[VULNERABLE] RCE possible at: http://52.63.41.222/wp-login.php with payload: system("ls")
wp-login.php	[VULNERABLE] RCE possible at: http://52.63.41.222/wp-login.php with payload: system("cat /etc/passwd")
wp-content/debug.log	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-content/debug.log with payload:

Figure-17: Vulnerability Scan Report

2.7 Integrating Chatbot into Website

For this task we have chosen Automotive industry i.e WaterTunnel car wash company. We will also evaluate organization's benefits, stakeholder's overview and customer's reaction upon using chatbot. This task will develop a comprehensive strategy for security leaders to effectively integrate chatbots into their organization's system along with managing associated risks and ethical issues. In this task we will also discuss about the problems on lack of clear guidance for IT leaders and other technical and non-technical staff on how to use and implement chatbots into their organization's security system and how could we get benefited from such type of GenAI in long runs. Many organizations are struggling to balance the potential benefits of using chatbots as customer service provider such as increased productivity, cost effective, consistency and availability. Apart from that we must be careful about the implementation strategies, ethical and security concerns, skills gap and customers adaptations.

The figure below shows the flow charts of customer service chatbots working process.

Figure-18: Flowchart of Chatbots working process

The flowchart shows when the customers send an enquiry message, the chatbot reads the text message and replies to the message that shows the available services in the format of “Hello! How can I help you today”? If the customers ask about price, then chatbots reply “our basic car wash package starts from 10$” and so on. If the customers’ requirement does not meet the criteria, then ask a feedback question, did you mean by this type of car wash (wash menu such as exterior wash, interior wash or both etc)? If the customer says yes, send feedback and if the customer says no then inform them the service is about to terminate and send them a “Thank you” message at the end.

here we have designed the SWOT analysis of customer service chatbot used in automobile survice industry.

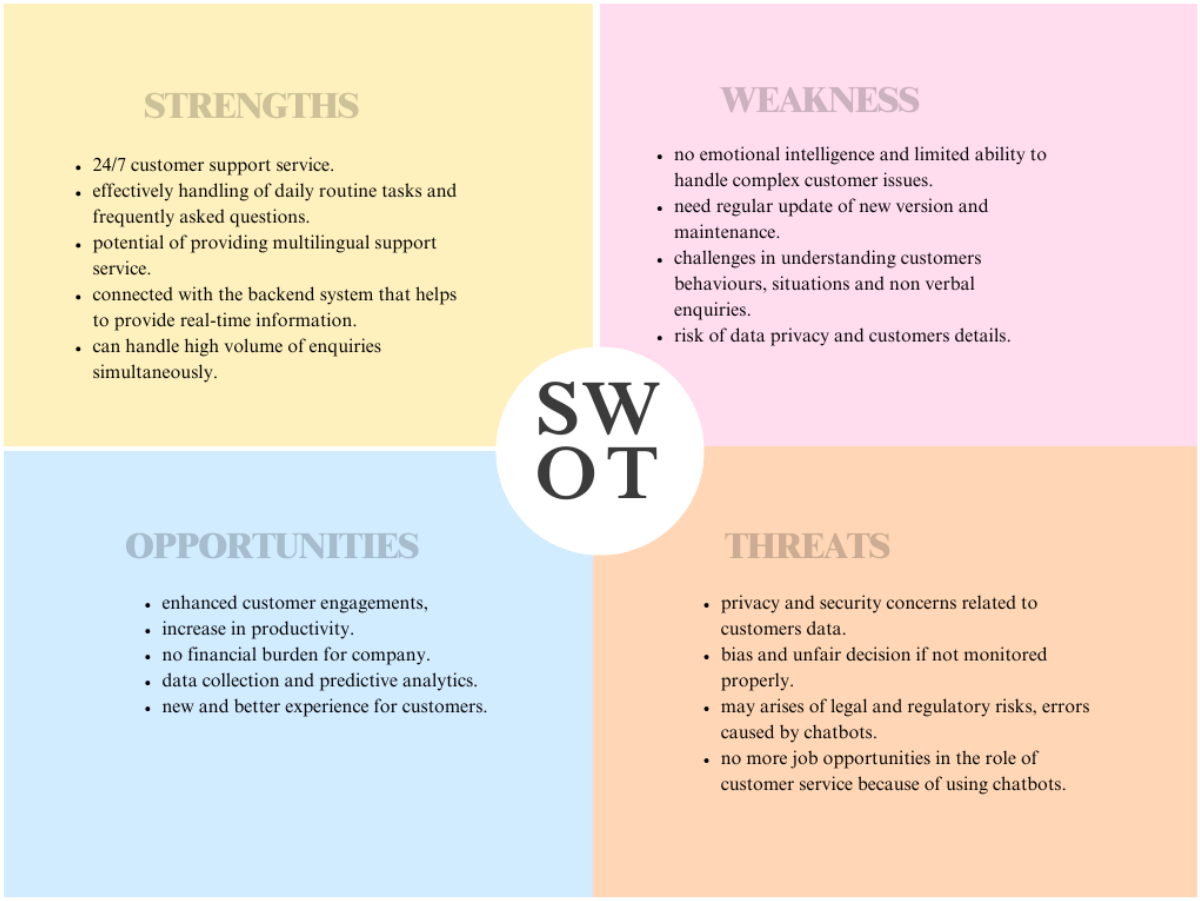


Figure-19: SWOT analysis of customer service chatbots used on Automotive Industry.

2.7.1 Ethical AI framework for chatbot

Industry such as WaterTunnel Car Wash within the automobile service have benefited from the integration of AI customer service chatbots, and there is a need to establish the ethical standard that is fit for use. This framework is intended to help cover such significant ethical requirements as protection of the user's identity and building trust between users.

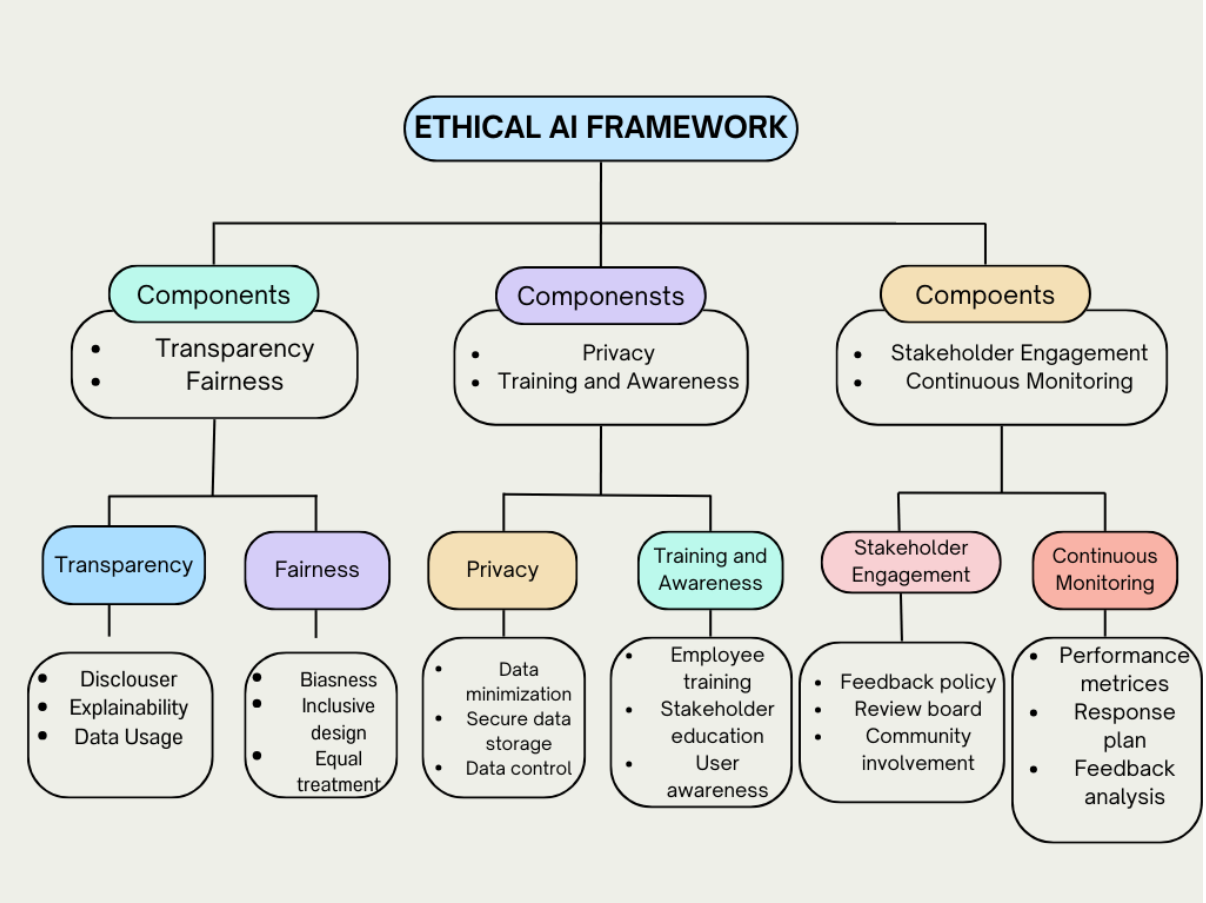


Figure-20: Ethical AI Framework for AI tool Chatbots in Automotive Industry

2.7.2 Chatbot Deployment and Testing

While integrating chatbot into the website we have made and named chatbot through FastBots.ai and import data from website. When integrating designed chatbot into our website we used embed script provided by FastBots as simple as we used plugin for easier integration. Then we monitor interaction, gather insights and update the chatbots for proper functioning:

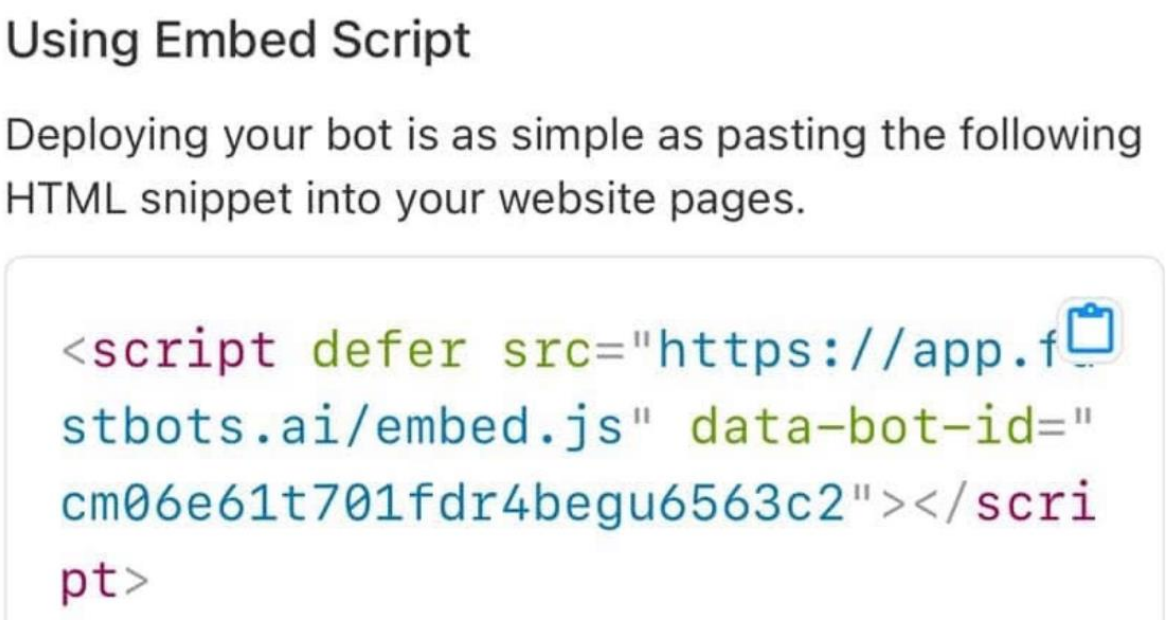


Figure-21: Embed script to link chatbot into website.

The figure below shows the proper functioning of chatbot and provide all the information belongs to the website.

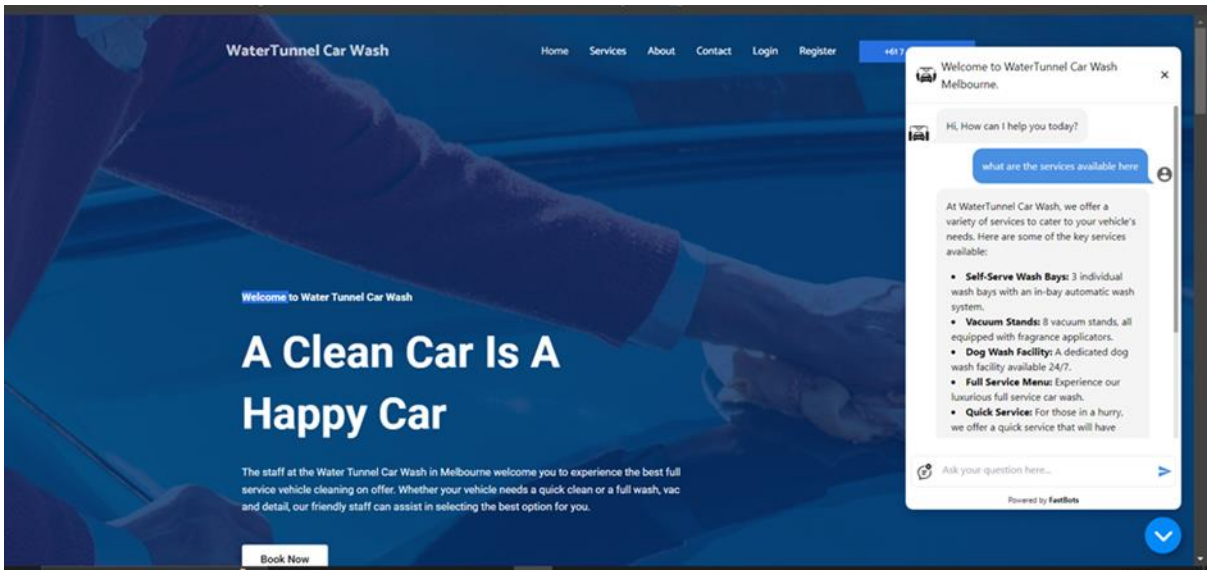


Figure-22: Integration of chatbot into website

2.8 List of Issues and Challenges and Mitigation

Issues and Challenges	**Mitigation**
Finding official WordPress.	We went through most of AMI in AWS most of them were paid, we chose WordPress certified by Bitnami because of its cost-effectiveness.
Selecting security group configuration while launching.	We were confused about whether to select SSH, HTTP or HTTPS. Later, we did google research and found out to select all three.
Finding WordPress admin password.	We waited up to 1 hr after completing launch in EC2 to receive username and password.
Unexpected cost.	The cost for launching WordPress will be deducted per hour rate so we must visit AWS web server regular.
While creating the Gantt chart, errors indicated by red lines appeared on the start and end dates of some tasks	We resolved this issue by adjusting the predecessors and re-entering the data, which corrected the errors and allowed us to generate the Gantt chart successfully.
It was difficult to select an appropriate industry and identify the relevant stakeholders.	We chose the automotive industry, specifically a water tunnel car wash. With the help of the manager, we gathered information about the stakeholders.
Scanning urls or multiple tests cases network latency can cause timeouts.	Use reasonable timeout values for ‘requests.get’ and ‘requests.post’ that we have used.
Automated vulnerability scanners can produce false positives.	Refined payloads for SQL injection and XSS have been used.
Using an anomaly detection model can yield false positives.	Training data for the anomaly detection model is representative of both normal and abnormal responses.
Training data for the anomaly detection model is	Our program includes memory-efficient data structures.

representative of both normal and abnormal responses.	
Pre-trained CodeBERT model might not always produce accurate result.	Fine-tune the CodeBERT model with a dataset of labelled vulnerability reports specific to our use case.

3. Delivered Technical Artifacts

Name	File	Description	PDF?
Risk Assessment Report	group11-risk-assessment-report.docx	Reports the process, assumptions and conclusions of the risk assessment.	Yes
Risk Assessment Table	group11-risk-assessment-table.xlsx	Details of the risks and security controls identified from the risk assessment.	Yes
AWS setup instruction	group11-AWS-setup-instructions.docx	Instructions for deploying the webserver in AWS.	Yes
Vulnerability Scan Test Script	group11-test-script.py	Python code used for vulnerability test, anomalies detection and vulnerability report.	No

4. Contribution Table

Student Name	Percent	Summary of Contribution	Technical Lead on Artefacts
Bhuwan Thapa	25%	- Deploy WordPress into AWS cloud server and create website. - Chatbot design. - System diagram. - Flow chart design. - Network design. - Design of Network/Security architecture.	- AWS setup. - Diagram design. - AI tool performance test. - Vulnerability scan report analysis.
Pratik Singh Dhami	25%	- Program coding for AI. - Chatbot integration. - Helping in WordPress deployment, agile diagram.	- AI tool developer. - Chatbot deployment. - Website test script.
Basanta Adhikari	25%	- Website design. - Gantt chart, kanban board design.	- Website designer. - Risk assessment table.

		- Project network diagram. - Risk Assessment. - Stakeholder Engagement matrix -Disaster Recovery Plan	- Risk assessment report.
Kirann Bhusal	25%	- AI tool developer. - Website design. - Helping in Network/System architecture and WordPress deployment in AWS. - WordPress GitHub Integration - Ethical framework design.	- Software developer. -Website Design -WordPress Setup -Network Architecture - Ethical framework design.

Conclusion

This project aimed to address the challenges and opportunities presented by generative AI in cybersecurity, focusing on balancing current skepticism with long-term potential. The primary goal was to develop a comprehensive strategy for security leaders within the automobile service industry to effectively integrate GenAI into their cybersecurity practices while managing associated risks and ethical issues. The project identified several key challenges within the industry, including the lack of clear implementation strategies, ethical and security concerns, skills gaps, risk assessment and mitigation needs, regulatory compliance uncertainties, and difficulties in convincing stakeholders of GenAI's long-term value. To address these challenges, the project proposed several solutions:

1. Recommendations for active collaboration between security leaders and business stakeholders of automobile service industry to ensure ethical and secure use of GenAI.
2. Created a framework for GenAI implementation in cybersecurity, including ethical principles, governance structures, risk assessment, data management, model development, transparency, human-AI collaboration, and continuous improvement.
3. Development of a risk assessment and mitigation plan for applying GenAI in cybersecurity, including cyber security risk assessment, threat and vulnerability analysis, and risk mitigation strategies.
4. Design of collaborative training and awareness programs for both technical and non-technical staff within the industry to bridge the skills gap and promote understanding of GenAI's capabilities and limitations.
5. Implementation of GenAI tools, including the development of chatbot integrated with website and vulnerability detection system, integrated into a website hosted on an AWS web server.

The project also addressed specific cybersecurity issues related to the implementation of GenAI into the system of automobile service industry, such as data protection and privacy, web application security, and potential for social engineering and phishing attacks. To mitigate these risks, the project suggested various security measures, including enhanced password

protection, vulnerability scanning, threat hunting queries, and the use of AWS security services.

Overall, this project provides valuable insights and practical strategies for organizations looking to implement GenAI in their cybersecurity practices. By addressing both short-term skepticism and long-term promise, it offers a balanced approach to navigating the rapidly evolving landscape of generative AI in cybersecurity.

References

Adamopoulou, E. and Moussiades, L., 2020. Chatbots: History, technology, and applications. *Machine Learning with applications*, *2*, p.100006.

Ahmad, N. and Alsmadi, I., 2021. Machine learning approaches for IoT security: A systematic literature review. Internet of Things, 14, p.100365

Al Fikri, M., Putra, F.A., Suryanto, Y. and Ramli, K., 2019. Risk assessment using NIST SP 800-30 revision 1 and ISO 27005 combination technique in profit-based organization: Case study of ZZZ information system application in ABC agency. *Procedia Computer Science*, *161*, pp.1206-1215.

Alshamrani, S. Myneni, A. Chowdhary and D. Huang, "A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1851-1877

Alwahedi, F., Aldhaheri, A., Ferrag, M.A., Battah, A. and Tihanyi, N., 2024. Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. Internet of Things and Cyber-Physical Systems, 10, p.100016.

Appiah, V., Nti, I.K. and Nyarko-Boateng, O., 2017. Investigating Websites and Web Application Vulnerabilities: Webmaster s Perspective. International Journal of Applied Information Systems, 12(3), pp.10-15.

ARS STAFF 2016, *Information-Technology*, viewed 22 August 2024, https://arstechnica.com/information-technology/2016/05/1b-bangladesh-heist-officials-say-swift-technicians-left-bank-vulnerable/

Asaad, R.R. and Saeed, V.A., 2022. A Cyber Security Threats, Vulnerability, Challenges and Proposed Solution. *Applied computing Journal*, pp.227-244.

AustLii, *Viewdb*, viewed 19 August 2024, (https://www8.austlii.edu.au/cgi-bin/viewdb/au/legis/cth/consol_act/taaa1979410/)

Bakare, S.S., Adeniyi, A.O., Akpuokwe, C.U. and Eneh, N.E., 2024. Data privacy laws and compliance: a comparative review of the EU GDPR and USA regulations. *Computer Science & IT Research Journal*, *5*(3), pp.528-543.

Balasubramaniam, N., Kauppinen, M., Hiekkanen, K. and Kujala, S., 2022, March. Transparency and explainability of AI systems: ethical guidelines in practice. In *International working conference on requirements engineering: foundation for software quality* (pp. 3-18). Cham: Springer International Publishing.

Blank, R. and Gallagher, P., 2012. Nist special publication 800-30 revision 1 guide for conducting risk assessments. *National Institute of Standards and Technology*.

Boiko, A. and Shendryk, V., 2017. System integration and security of information systems. Procedia Computer Science, 104, pp.35-42.

Bommasani, R. et al. (2021) 'On the Opportunities and Risks of Foundation Models', arXiv:2104.12547.

Bruschi, D. and Diomede, N. (2022) 'A framework for assessing AI ethics with applications to cybersecurity', AI and Ethics, 2, pp. 1005-1017.

Canadian Centre for Cyber Security (2023) Generative Artificial Intelligence. Available at: https://www.cyber.gc.ca/en/guidance/generative-artificial-intelligence-ai-itsap00041 (Accessed: 19 August 2024).

Chandrasekaran, A.S., 2024. Harnessing the Power of Generative Artificial Intelligence (GenAI) in Governance, Risk Management, and Compliance (GRC).

Frederiksen, J.R. and White, B.Y., 1998. Teaching and learning generic modeling and reasoning skills. Interactive Learning Environments, 5(1), pp.33-51.

Chui, M., Hazan, E., Roberts, R., Singla, A. and Smaje, K., 2023. The economic potential of generative AI.

Cybersecurity and Infrastructure Security Agency (2023) CISA Roadmap for Artificial Intelligence. Available at: https://www.cisa.gov/sites/default/files/2023-11/2023-2024_CISA-Roadmap-for-AI_508c.pdf (Accessed: 20 August 2024).

Dhoni, P. and Kumar, R., 2023. Synergizing generative ai and cybersecurity: Roles of generative ai entities, companies, agencies, and government in enhancing cybersecurity. *Authorea Preprints*.

Dhoni, P. and Kumar, R., 2023. Synergizing generative ai and cybersecurity: Roles of generative ai entities, companies, agencies, and government in enhancing cybersecurity. *Authorea Preprints*.

Dimakopoulou, A. and Rantos, K., 2024. Comprehensive Analysis of Maritime Cybersecurity Landscape Based on the NIST CSF v2. 0. *Journal of Marine Science and Engineering*, *12*(6), p.919.

Ding, D. et al. (2022) 'A Survey on Security, Privacy and Fairness of Machine Learning in Computer Vision', arXiv:2212.08073.

Dwivedi, Y.K., Hughes, L., Cheung, C.M., Coombs, C., Doyle, R., Dubey, R., Eirug, A., Grover, V., Gupta, B., Gustafsson, A. and Hinsch, C., 2024. Generative AI (GenAI) – Opening Pandora's box: Towards a research agenda focusing on darker sides, challenges, implications and future prospects. International Journal of Information Management, 75, p.102742

Elder, J. & Elder, S. 2019. 'Faster Disaster Recovery The Business Owner's Guide to Developing a Business Continuity Plan'. Hoboken, New Jersey, USA. John Wiley & Sons Inc. E-book, viewed 23 August 2024, https://ebookcentral.proquest.com/lib/xamkebooks/reader.action?docID=5741229&query=Business+Continuity

European Court of Auditors (2023) ECA AI Strategy 2024-2025. Available at: https://www.eca.europa.eu/ECAPublications/ECA-AI-Strategy-2024-2025/ECA-AI-Strategy-2024-2025_EN.pdf (Accessed: 23 August 2024).

European Parliament (2023) EU AI Act: first regulation on artificial intelligence. Available at: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (Accessed: 24 August 2024).

Evan and Shimon 2015, *Politics*, viewed 21 August 2024, https://edition.cnn.com/2015/04/07/politics/how-russians-hacked-the-wh/index.html

Federal Trade Commission 2024, *Enforcement*, viewed 24 August 2024, https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement

Forbes Technology Council (2024) 'Generative AI's Challenges To Cybersecurity', Forbes, 19 April. Available at: https://www.forbes.com/councils/forbestechcouncil/2024/04/19/generative-ais-challenges-to-cybersecurity/ (Accessed: 25 August 2024).

Forbes Technology Council (2024) 'The Ethics Of AI: Balancing Innovation With Responsibility', Forbes, 8 February. Available at: https://www.forbes.com/councils/forbestechcouncil/2024/02/08/the-ethics-of-ai-balancing-innovation-with-responsibility/ (Accessed: 22 August 2024).

Furnell, S. and Clarke, N., 2012. Power to the people? The evolving recognition of human aspects of security. Computers & Security, 31(8), pp.983-988.

Gallery, C., Retail innovation 2: The future of online selling. In *Fashion Business and Digital Transformation* (pp. 216-254). Routledge.

Granjal, J., Monteiro, E. and Silva, J.S., 2013, May. End-to-end transport-layer security for Internet-integrated sensing applications with mutual and delegated ECC public-key authentication. In *2013 IFIP Networking Conference* (pp. 1-9). IEEE.

Gupta, M., Akiri, C., Aryal, K., Parker, E. and Praharaj, L., 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*.

Habbal, A., Ali, M.K. and Abuzaraida, M.A., 2024. Artificial Intelligence Trust, risk and security management (AI trism): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications*, *240*, p.122442.

Hasal, M., Nowaková, J., Ahmed Saghair, K., Abdulla, H., Snášel, V. and Ogiela, L., 2021. Chatbots: Security, privacy, data protection, and social aspects. *Concurrency and Computation: Practice and Experience*, *33*(19), p.e6426.

Huang, K., Ponnapalli, J., Tantsura, J. and Shin, K.T., 2024. Navigating the GenAI Security Landscape. In *Generative AI Security: Theories and Practices* (pp. 31-58). Cham: Springer Nature Switzerland.

Huang, K., Yeoh, J., Wright, S. and Wang, H., 2024. Build your security program for genai. In *Generative AI Security: Theories and Practices* (pp. 99-132). Cham: Springer Nature Switzerland.

Industry, *regulations-and-standards*, viewed 19 August 2024, https://www.industry.gov.au/regulations-and-standards.

Katsumata, P., Hemenway, J. and Gavins, W., 2010, October. Cybersecurity risk management. In *2010-MILCOM 2010 Military Communications Conference* (pp. 890-895). IEEE.

Kaur, R., Gabrijelčič, D. and Klobučar, T., 2023. Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, p.101804.

Kearns, M. and Roth, A. (2020) 'Ethical algorithm design should guide technology regulation', Harvard Data Science Review, 2(1).

Krishnamurthy, O., 2023. Enhancing Cyber Security Enhancement Through Generative AI. *International Journal of Universal Science and Engineering*, *9*, pp.35-50.

Lakhno, V., Blozva, A., Kasatkin, D., Chubaievskyi, V., Shestak, Y., Tyshchenko, D. and Brzhanov, R., 2022. Experimental studies of the features of using waf to protect internal services in the zero trust structure. *J Theor Appl Inf Technol*, *100*(3), pp.705-721.

Legislation 2024, *Home -Acts,* viewed 19 August 2024, https://www.legislation.gov.au/C2004A03712/latest/text

Limaj, B. (2023) 'Ethical Considerations in AI-Powered Cybersecurity', Medium, 15 May. Available at: https://medium.com/@besniklimaj/ethical-considerations-in-ai-powered-cybersecurity-45cd83db90e0 (Accessed: 24 August 2024).

Lucchi, N., 2023. ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, pp.1-23.

Mad Devs (2023) Artificial Intelligence in Cybersecurity. Available at: https://maddevs.io/blog/artificial-intelligence-in-cybersecurity/ (Accessed: 20 August 2024).

McKay, K. and Cooper, D., 2017. *Guidelines for the selection, configuration, and use of transport layer security (TLS) implementations* (No. NIST Special Publication (SP) 800-52 Rev. 2 (Draft)). National Institute of Standards and Technology.

Microsoft (2023) Microsoft's AI safety policies. Available at: https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/ (Accessed: 13 August 2024).

Möller, D.P., 2023. NIST cybersecurity framework and MITRE cybersecurity criteria. In *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices* (pp. 231-271). Cham: Springer Nature Switzerland.

National Institute of Standards and Technology (2023) AI Risk Management Framework. Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (Accessed: 25 August 2024).

OAIC, *consumer-data-right,* viewed 19 August 2024, https://www.oaic.gov.au/consumer-data-right/consumer-data-right-legislation,-regulation-and-definitions/consumer-data-right-legislation

OAIC, *Privacy,* viewed 19 August 2024, https://www.oaic.gov.au/privacy/notifiable-data-breaches/about-the-notifiable-data-breaches-scheme**.**

OpenAI (2023) GPT-4 Technical Report. Available at: https://cdn.openai.com/papers/gpt-4.pdf (Accessed: 19 August 2024).

Patibandla, K.R., 2024. Design and Create VPC in AWS. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, *1*(1), pp.273-282.

Rajapaksha, S., Senanayake, J., Kalutarage, H. and Al-Kadri, M.O., 2022, December. Ai-powered vulnerability detection for secure source code development. In International Conference on Information Technology and Communications Security (pp. 275-288). Cham: Springer Nature Switzerland.

Rane, N., 2023. Role and challenges of ChatGPT and similar generative artificial intelligence in business management. *Available at SSRN 4603227*.

Ranimäki, M., 2023. Web security and hacking (Master's thesis, Itä-Suomen yliopisto).

Safitri, E.H.N. and Kabetta, H., 2023, August. Cyber-Risk Management Planning Using NIST CSF V1. 1, ISO/IEC 27005: 2018, and NIST SP 800-53 Revision 5 (A Study Case to ABC Organization). In *2023 IEEE International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs)* (pp. 332-338). IEEE.

Schmoker, M., 1999. Results: The key to continuous school improvement. ASCD.

Bada, M., Sasse, A.M. and Nurse, J.R., 2019. Cyber security awareness campaigns: Why do they fail to change behaviour?. arXiv preprint arXiv:1901.02672.

Snedaker, S., 2013. *Business continuity and disaster recovery planning for IT professionals*. Newnes, viewed 20 August 2024, https://books.google.com.au/books?hl=en&lr=&id=vT8TAAAAQBAJ&oi=fnd&pg=PP1&dq=cybersecurity+business+continuity+plan&ots=d2uAj9138-&sig=bu9fbT6qoOioIWEsrnlxEZouoh8#v=onepage&q&f=false

Thorgersen, S. and Silva, P.I., 2021. *Keycloak-identity and access management for modern applications: harness the power of Keycloak, OpenID Connect, and OAuth 2.0 protocols to secure applications*. Packt Publishing Ltd.

Townsend, R. 2022. *What is a Disaster Recovery Plan and Why is it Important? Nexstor. WWW document. Updated 31 August 2022*, viewed 21 August 2024, https://nexstor.com/what-is-disaster-recovery-plan/.

Wang, X., Lin, X. and Shao, B., 2022. How does artificial intelligence create business agility? Evidence from chatbots. *International journal of information management*, *66*, p.102535.

World Health Organization (2024) WHO releases AI ethics and governance guidance for large multi-modal models. Available at: https://www.who.int/news/item/18-01-2024-who-

releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models (Accessed: 25 August 2024).

Yaqoob, I., Ahmed, E., ur Rehman, M.H., Ahmed, A.I.A., Al-garadi, M.A., Imran, M. and Guizani, M., 2017. The rise of ransomware and emerging security challenges in the Internet of Things. *Computer Networks*, *129*, pp.444-458.

Central Queensland university (CQU) Intellectual property
COIT20265: NETWORK AND INFORMATION SECURITY PROJECT (HT2, 2024)

RISK ASSESSMENT REPORT

PROJECT:
GENERATIVE AI: NAVIGATING SHORT-TERM SKEPTISM AND LONG-TERM PROMISE

UNDER THE ESTEEMED GUIDANCE OF
UNIT CO-ORDINATOR: FARIZA SABRINA
MENTOR: DR. AHMEDI AZRA

TEAM MEMBERS:
BASANTA ADHIKARI (12211752)
BHUWAN THAPA (12196590)
KIRAN BHUSAL (12211570)
PRATIK SHINGH DHAMI (12209929)

Table of Contents

1. Risk Assessment and Mitigation Plan

1.1 Cyber Security Risk Assessment

Risk Assessment of GenAI application in cybersecurity for Automobile Service Industry is conducted following NIST Special Publication 800-30 Revision 1, titled "Guide for Conducting Risk Assessments," which offers a structured framework for evaluating risks to information systems (Blank and Gallagher 2012). The guide emphasizes a systematic approach starting with the preparation phase, where organizations define the methodology of the risk assessment. This involves identifying the information system boundaries and assets at risk (Fikri et al. 2019). During the assessment phase, identify and analyse potential threats and vulnerabilities, assessing their likelihood and impact to determine overall risk levels. The process includes identifying threats such as cyber-attacks or insider threats and vulnerabilities like unpatched software or misconfigured systems. (Safitri and Kabetta 2023). The guide suggests categorizing likelihood and impact to evaluate the risk comprehensively.

The next steps involve selecting and implementing appropriate controls to mitigate identified risks. NIST CSF 2.0 core functions provides the framework which is used as cybersecurity controls to mitigate Risks. The CSF Core Functions GOVERN, IDENTIFY, PROTECT, DETECT, RESPOND, and RECOVER organize cybersecurity outcomes at their highest level. The NIST Cybersecurity Framework (CSF) outlines six core functions designed to help organizations manage and respond to cybersecurity risks comprehensively. The first function, **Govern (GV)**, focuses on establishing and communicating the organization's cybersecurity risk management strategy, policies, and governance structures. This function ensures that roles and responsibilities are clearly defined and that cybersecurity efforts align with the organization's broader enterprise risk management (ERM) strategy.

Next, the **Identify (ID)** function helps organizations gain a clear understanding of their current cybersecurity risks by identifying critical assets such as data, hardware, software, systems, and people. By understanding these assets and their associated risks, organizations can prioritize their cybersecurity efforts to address the most pressing vulnerabilities.

The **Protect (PR)** function involves implementing safeguards to manage the identified risks, helping to prevent or reduce the likelihood of cybersecurity incidents. These safeguards include measures like access control, identity management, and data security, along with training programs to raise awareness among employees.

The **Detect (DE)** function ensures that organizations have the capabilities to identify and analyze potential cybersecurity threats or incidents in a timely manner. By monitoring for

anomalies and indicators of compromise, organizations can quickly detect attacks before they cause significant damage.

The **Respond (RS)** function comes into play, guiding organizations in taking action to contain and mitigate the impact of cybersecurity incidents. This function involves conducting detailed incident analysis, managing the containment process, and ensuring that the incident is communicated appropriately to stakeholders. The goal is to minimize the damage caused by the incident and to take steps to prevent its recurrence.

Finally, the **Recover (RC)** function focuses on restoring affected assets and operations following a cybersecurity incident. Recovery efforts aim to quickly return to normal business operations while reducing the long-term impact of the event. This function also emphasizes clear communication during the recovery process to ensure that all stakeholders are informed, and that the organization can continue to operate smoothly in the aftermath of an incident.

These controls can be technical, administrative, or physical and should be continuously monitored to ensure their effectiveness. Regular reviews and adjustments are crucial as the risk environment evolves (Möller, 2023). Overall, NIST SP 800-30 Rev. 1 and NIST CSF 2.0 provides a detailed methodology for identifying, analysing, and managing risks, helping organizations protect their information systems and data effectively. In cybersecurity, understanding the relationships between threats, vulnerabilities, and assets is essential for developing robust protection strategies. The Threat-Vulnerability-Asset (TVA) model helps in identifying and managing these elements systematically.

Threats are potential events or actions that can exploit vulnerabilities to cause harm or damage to an organization. They represent the external or internal forces that pose risks. Examples include cyberattacks, insider threats, and natural disasters. (Asaad and Saeed 2022)

Vulnerabilities are weaknesses or flaws within a system that can be exploited by threats. They represent gaps in security that could be used to breach or damage the system. Vulnerabilities can be due to software bugs, misconfigurations, lack of updates, or weak authentication mechanisms. (Asaad and Saeed 2022)

Assets are the resources or components of an organization that need protection. They can include hardware, software, data, and processes. Assets are valuable because they store, process, or transmit sensitive information or are crucial for operations. (Asaad and Saeed 2022)

1.2 TVA in Cybersecurity:

Threat-Vulnerability-Asset Relationships:

1. Threat: A cyberattack aiming to exploit a weakness.

2. Vulnerability: The weakness in the system, such as unpatched software.
3. Asset: The system or data that the threat is targeting.

For example, if a threat is a malware attack, the vulnerability could be an unpatched software, and the asset could be a server storing critical data.

Using TVA for Risk Management:

1. Identify Threats: Understand what potential threats exist (e.g., phishing attacks, data breaches).
2. Assess Vulnerabilities: Identify weaknesses that could be exploited by these threats (e.g., weak passwords, outdated software).
3. Protect Assets: Focus on securing assets that are at risk (e.g., sensitive data, critical systems).

TVA Example:

1. Threat: Software attack (e.g., code injection).
2. Vulnerability: Poorly secured API endpoints.
3. Asset: GenAI application that processes and stores customer data.

In this case, the threat of a software attack could exploit the vulnerability of poorly secured API endpoints to compromise the asset, which is the GenAI application. This would lead to potential data breaches and operational disruption.

Understanding these relationships, risk analysis is conduct and organizations can prioritize security efforts to address the most critical vulnerabilities and protect their valuable assets from potential threats. This systematic approach helps in developing effective risk management and mitigation strategies (Blank and Gallagher 2012). Table below shows the Assets which are associated with the organization.

1.3 Assets Table

Asset No.	Asset Type	Asset
1.	**Hardware**	Servers (Cloud-based) that host the GenAI application and website
		Web Server for hosting the company's website
		Routers for network connectivity
		Firewalls to protect the network
		Switches to connect network devices
		Workstations/Computers used by staff
		Telecommunications Equipment (phones, intercoms)
		Load Balancers (Cloud-based) to manage website traffic
		Point of Sale (POS) Systems for handling customer payments
		Carwash Equipment (automated carwash machines, sensors, etc.)
2.	**Software**	GenAI Application
		Content Management System (CMS) for website updates
		Operating Systems on servers and workstations
		Security Tools (e.g., antivirus, threat detection software)
		APIs for integrating GenAI with the website and other systems
		SSL/TLS Certificates for secure web communication
		Point of Sale (POS) Software
		Customer Relationship Management (CRM) Software
		Business Management Software (inventory, scheduling, etc.)
		Monitoring Tools for network and application performance
		Logging Systems for tracking activities and troubleshooting
3.	**Data**	Training Data for the GenAI model
		Customer Data (personal details, service history, payment info)
		Operational Data (logs, configurations, business metrics)
		Proprietary Information (business strategies, pricing models)
		Financial Data (transaction records, financial statements)
		Compliance Documentation Data (data protection, safety regulations)

		Legal Agreements Data (customer terms, supplier contracts, privacy policies)
4.	**Process**	Network Setup Process
		Device Configuration Process
		GenAI Tool Development and Deployment Process
		Customer Interaction Process (in-person and online)
		Service Booking Process (online and on-site)
		Payment Processing Process
		Change Management Process
		Incident Response Process
		Data Management Process
		Access Control Process for IT systems and physical premises
		Employee Training Process (for using IT systems and carwash equipment)
		Disaster Recovery Plan for IT infrastructure and carwash operations process
5.	**Cloud Services**	Cloud Storage for data backups
		Cloud Compute Instances for running the GenAI model
		Cloud Security Services to protect cloud-based resources

Table-1 Assets Table

1.4 Threat Table

Threat No	Threat Name	Description	Related Assets
1	**Software Attacks**	Includes Malware, DoS (Denial of Service), DDoS (Distributed Denial of Service), MITM (Man in the Middle), Sniffing, DNS Poisoning, Spoofing, Code Injection, Backdoors, and Ransomware.	GenAI Application, Operating System, Security Tools, APIs, Servers (Cloud-based)
2	**Human Errors**	Refers to Phishing, Server Misconfiguration, Social	All Assets (Hardware, Software, Data, Processes)

		Engineering, and Accidental Data Deletion.	
3	**Trespass**	Unauthorized Access via password attacks, such as Dictionary Attack, Brute Force Attack, Shoulder Surfing, and Physical Tampering.	Servers, Router, Switch, Computer, Customer Data, Access Control Process, APIs
4	**Information Extortion**	Ransomware attacks leading to Blackmail or Information Disclosure for Financial Gain.	Customer Data, Proprietary Information, Operational Data, GenAI Application
5	**Hardware Failures/Errors**	Includes Drive Failures, Server Crashes, and Network Hardware Malfunctions.	Servers, Router, Firewall, Switch, Computer
6	**Wi-Fi Eavesdropping**	Capturing Sensitive Information, Passwords, or Other Confidential Data through Unauthorized Wireless Network Monitoring.	Router, Switch, Access Control Process, Customer Data
7	**Software Failures/Errors**	Bugs, Code Performance Issues, Loopholes, and Unpatched Vulnerabilities.	GenAI Application, Operating System, APIs, Security Tools, Chatbot Development Process, Service Booking Process
8	**Sabotage and Vandalism**	Deliberate Destruction or Tampering of Assets, including Physical Damage to Hardware and Digital Destruction (e.g., Data Wiping, Defacement).	Servers, Firewall, Proprietary Information, Data Management Process
9	**Forces of Nature**	Natural Disasters such as Cyclones, Fires, Floods, and Lightning Strikes, leading to Data Loss, Hardware Damage, or Operational Downtime.	All Physical Hardware (Servers, Router, Firewall, Switch, Computer), Training Data, Customer Data

10	**Data Breaches**	Unauthorized access to sensitive data, resulting in information leakage and compliance violations.	Customer Data, Proprietary Information, Training Data, Operational Data
11	**Insider Threats**	Malicious actions taken by employees or contractors, such as data theft, sabotage, or misuse of access privileges.	All Assets (Hardware, Software, Data, Processes)
12	**Supply Chain Attacks**	Attacks targeting the organization's suppliers, leading to compromised components or software used within the infrastructure.	GenAI Application, APIs, Security Tools, Operating System
13	**API Abuse**	Unauthorized use or exploitation of API endpoints, leading to data breaches, service disruption, or unauthorized actions.	APIs, GenAI Application, Operating System
14	**Credential Theft**	Theft of authentication credentials through phishing, malware, or brute-force attacks, leading to unauthorized access.	Servers, Router, Firewall, Switch, GenAI Application, Customer Data, Access Control Process
15	**Cloud Security Misconfigurations**	Improper configuration of cloud services, leading to exposure of sensitive data or unauthorized access to cloud-based systems.	Servers (Cloud-based), GenAI Application, Customer Data, Proprietary Information
16	**Zero-Day Exploits**	Exploitation of unknown vulnerabilities in software before they can be patched, leading to unauthorized access or system compromise.	Operating System, GenAI Application, APIs, Security Tools
17	**Regulatory Non-Compliance**	Failure to comply with relevant data protection and cybersecurity	All Data (Customer Data, Training Data, Operational Data,

		regulations, leading to fines, legal action, and reputational damage.	Proprietary Information), Processes related to Data Management and Compliance
18	**Data Poisoning**	Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs.	GenAI Application, Training Data
19	**Model Inversion**	Attacker's reverse-engineering the model to infer sensitive information from the training data.	GenAI Application, Training Data
20	**Adversarial Attacks**	Small, crafted inputs causing the GenAI model to make incorrect or harmful predictions.	GenAI Application
21	**Model Drift**	Performance degradation over time as the model encounters data differing from its training set.	GenAI Application
22	**Hallucination**	Generation of nonsensical or incorrect outputs that appear plausible.	GenAI Application
23	**Bias and Discrimination**	Inherited biases from training data leading to unfair or discriminatory outputs.	GenAI Application
24	**Explain ability**	Difficulty in understanding and interpreting model outputs, obscuring errors and reducing trust.	GenAI Application
25	**Dependency on Training Data Quality**	Inadequate or biased training data leading to flawed models and erroneous outputs.	GenAI Application
26	**Intellectual Property Violations**	Unintentional plagiarism or copyright infringement due to the replication of protected content.	GenAI Application, Proprietary Information

27	**Overfitting**	The model performs well on training data but poorly on new, unseen data.	GenAI Application
28	**Malicious Use of GenAI**	GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks.	GenAI Application
29	**Privacy Invasion**	The application inadvertently exposing sensitive personal or organizational data through its outputs.	GenAI Application, Customer Data
30	**Algorithmic Manipulation**	Attackers manipulating the GenAI's algorithm to produce biased or harmful outputs.	GenAI Application
31	**Unauthorized Access**	Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs.	GenAI Application
32	**Ethical Violations**	GenAI producing outputs that violate ethical standards, such as generating offensive content.	GenAI Application

Table-2 Threats Table

1.5 Vulnerability

Vulnerabilities are weaknesses or flaws within a system that can be exploited by threats to cause harm or unauthorized access. In the context of Implementing GenAI application for the Automobile Service Industry several critical vulnerabilities have been identified and are in table below.

1.5.1 Vulnerability Table

Threat	Asset	Vul.No	TVA	Vulnerability

Software Attacks	Firewall	1	T1V1A1	Weak authentication, such as easily guessable passwords.
Software Attacks	GenAI Application	2	T1V2A2	Unpatched vulnerabilities in the GenAI codebase.
Software Attacks	APIs	3	T1V3A3	Poorly secured API endpoints prone to injection attacks.
Human Errors	Server (Cloud-based)	4	T2V4A1	Misconfiguration during server setup.
Human Errors	Operating System	5	T2V5A2	Lack of timely updates and patches.
Trespass	Router	6	T3V6A1	Default credentials not changed, leading to unauthorized access.
Information Extortion	Customer Data	7	T4V7A3	Lack of encryption, leading to data exposure during extortion.
Hardware Failures/Errors	Servers (Cloud-based)	8	T5V8A1	Single point of failure in server hardware.
Wi-Fi Eavesdropping	Switch	9	T6V9A1	Insufficient encryption on wireless communication.
Software Failures/Errors	GenAI Application	10	T7V10A2	Bugs and code performance issues leading to system crashes.
Sabotage and Vandalism	Servers (Cloud-based)	11	T8V11A1	Lack of physical security, leading to tampering.
Forces of Nature	All Physical Hardware	12	T9V12A1	Lack of disaster recovery plans for natural disasters.
Data Breaches	Customer Data	13	T10V13A3	Weak access controls leading to unauthorized data access.

Insider Threats	All Assets	14	T11V14A1-5	Insufficient monitoring and auditing of employee activities.
Supply Chain Attacks	GenAI Application	15	T12V15A2	Compromised third-party software or components used within the application.
API Abuse	APIs	16	T13V16A2	Insufficient rate limiting and authentication mechanisms on APIs.
Credential Theft	Access Control Process	17	T14V17A4	Weak password policies leading to credential theft.
Cloud Security Misconfigurations	Servers (Cloud-based)	18	T15V18A1	Improper configuration of cloud services, leading to exposure of sensitive data.
Zero-Day Exploits	Operating System	19	T16V19A2	Unpatched zero-day vulnerabilities exploited by attackers.
Regulatory Non-Compliance	Customer Data, Training Data, Proprietary Information	20	T17V20A3	Lack of adherence to data protection and cybersecurity regulations, leading to legal and financial penalties.
Data Poisoning	GenAI Application	21	T18V21A2	Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs.
Model Inversion	GenAI Application	22	T19V22A2	Attacker's reverse-engineering the model to infer sensitive information from the training data.
Adversarial Attacks	GenAI Application	23	T20V23A2	Small, crafted inputs causing the GenAI model to

				make incorrect or harmful predictions.
Model Drift	GenAI Application	24	T21V24A2	Performance degradation over time as the model encounters data differing from its training set.
Hallucination	GenAI Application	25	T22V25A2	Generation of nonsensical or incorrect outputs that appear plausible.
Bias and Discrimination	GenAI Application	26	T23V26A2	Inherited biases from training data leading to unfair or discriminatory outputs.
Dependency on Training Data Quality	GenAI Application	28	T24V27A2	Inadequate or biased training data leading to flawed models and erroneous outputs.
Intellectual Property Violations	Proprietary Information	29	T25V28A3	Unintentional plagiarism or copyright infringement due to the replication of protected content.
Overfitting	GenAI Application	30	T26V29A2	The model performs well on training data but poorly on new, unseen data.
Malicious Use of GenAI	GenAI Application	31	T27V30A2	GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks.
Privacy Invasion	GenAI Application and Data	32	T28V31A3	The application inadvertently exposing sensitive personal or

				organizational data through its outputs.
Algorithmic Manipulation	GenAI Application	33	T29V32A2	Attackers manipulating the GenAI's algorithm to produce biased or harmful outputs.
Unauthorized Access	All Assets	34	T30V33A1-5	Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs.
Ethical Violations	GenAI Application	35	T31V34A2	GenAI producing outputs that violate ethical standards, such as generating offensive content.

Table-3 Vulnerability Table

1.6 Risk Rank

In risk assessment, the risk rank is a crucial metric used to prioritize vulnerabilities based on their likelihood of exploitation and the potential impact if they are exploited. The likelihood measures how probable it is that a vulnerability will be exploited, categorized as High, Moderate, or Low. Impact evaluates the potential consequences of exploitation, also categorized as High, Moderate, or Low. By combining these two factors, the overall risk is determined (Blank and Gallagher 2012).

For example, a vulnerability with High likelihood and High impact is deemed a High-risk issue, requiring immediate attention and remediation. Conversely, a vulnerability with Low likelihood and Low impact is considered Low risk and is given a lower priority. The risk rank assigns a numerical value or category, such as Rank 1 for Critical High risks that demand urgent action, rank 2 for Moderate risks that should be addressed soon, and Rank 3 correspond to risks that are less critical and can be addressed as resources permit. This prioritization ensures that the most significant threats are managed first, effectively reducing the potential for substantial damage or disruption to the organization.

Likelihood	Impact	Combined Risk	Risk
High	Low	High /Low	Low
Moderate	Low	Moderate /Low	Low
Low	Low	Low /Low	Low
High	Moderate	High /Moderate	Moderate
Moderate	Moderate	Moderate /Moderate	Moderate
Low	Moderate	Low/Moderate	Low
High	High	High /High	High
Moderate	High	Moderate /High	Moderate
Low	High	Low /High	Low

Table-4 Risk Rank Value Table

1.7 Vulnerability occurrence and Risk AssessmentTable

Vul No.	Vulnerability	Likelihood	Impact	Risk	Rank
1	Weak authentication, such as easily guessable passwords.	Moderate	High	Moderate	2
2	Unpatched vulnerabilities in the GenAI codebase.	High	High	High	1
3	Poorly secured API endpoints prone to injection attacks.	High	Moderate	Moderate	2
4	Misconfiguration during server setup.	Moderate	High	Moderate	2
5	Lack of timely updates and patches.	Moderate	High	Moderate	2
6	Default credentials not changed, leading to unauthorized access.	High	High	High	1
7	Lack of encryption, leading to data exposure during extortion.	Moderate	High	Moderate	2
8	Single point of failure in server hardware.	Low	High	Moderate	3
9	Insufficient encryption on wireless communication.	Moderate	Moderate	Moderate	3
10	Bugs and code performance issues leading to system crashes.	High	High	High	1
11	Lack of physical security, leading to tampering.	Moderate	High	Moderate	2
12	Lack of disaster recovery plans for natural disasters.	Low	High	Moderate	3
13	Weak access controls leading to unauthorized data access.	High	High	High	1

14	Insufficient monitoring and auditing of employee activities.	Moderate	High	Moderate	2
15	Compromised third-party software or components used within the application.	Moderate	High	Moderate	2
16	Insufficient rate limiting and authentication mechanisms on APIs.	High	Moderate	High	2
17	Weak password policies leading to credential theft.	High	High	High	1
18	Improper configuration of cloud services, leading to exposure of sensitive data.	Moderate	High	Moderate	2
19	Unpatched zero-day vulnerabilities exploited by attackers.	Moderate	High	Moderate	2
20	Lack of adherence to data protection and cybersecurity regulations, leading to legal and financial penalties.	Moderate	High	Moderate	2
21	Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs.	High	High	High	1
22	Attackers reverse-engineering the model to infer sensitive information from the training data.	Moderate	High	Moderate	2
23	Small, crafted inputs causing the GenAI model to make	High	High	High	1

	incorrect or harmful predictions.				
24	Performance degradation over time as the model encounters data differing from its training set.	Moderate	Moderate	Moderate	3
25	Generation of nonsensical or incorrect outputs that appear plausible.	Moderate	Moderate	Moderate	3
26	Inherited biases from training data leading to unfair or discriminatory outputs.	Moderate	High	Moderate	3
27	Difficulty in understanding and interpreting model outputs, obscuring errors and reducing trust.	Moderate	Moderate	Moderate	3
28	Inadequate or biased training data leading to flawed models and erroneous outputs.	Moderate	High	Moderate	2
29	Unintentional plagiarism or copyright infringement due to the replication of protected content.	Moderate	High	Moderate	2
30	The model performs well on training data but poorly on new, unseen data.	Moderate	Moderate	Moderate	3
31	GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks.	High	High	High	1
32	The application inadvertently exposing sensitive personal or	Moderate	High	Moderate	2

	organizational data through its outputs.				
33	Attackers manipulating the GenAI's algorithm to produce biased or harmful outputs.	High	High	High	1
34	Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs.	High	High	High	1
35	GenAI producing outputs that violate ethical standards, such as generating offensive content.	Moderate	High	Moderate	2

Table-5 Vulnerability occurrence and Risk rank Table

1.8 Risk Mitigation Strategies:

To address the identified vulnerabilities and associated risks effectively, a comprehensive risk mitigation strategy is developed. According to the NIST Cybersecurity Framework (CSF) 2.0 To organize vulnerabilities and corresponding mitigation strategies, each vulnerability is categorized by NIST core functions (Identify, Protect, Detect, Respond, Recover and Govern) with clear mappings to the respective mitigation strategies (Dimakopoulou and Rantos 2024). This strategy includes preventative measures, detection mechanisms, and response plans tailored to the specific vulnerabilities and their risk rankings (Katsumata et Al. 2010), (Blank and Gallagher 2012). Robust mitigation strategies categorized by NIST CSF category and sub-category risk rank are as Follows:

1.8.1 High-Risk Vulnerabilities Mitigation

1. **Unpatched Vulnerabilities (Vul No. 2):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-02 (Software is maintained, replaced, and removed commensurate with risk).

Implement a robust patch management process. Regularly update and patch all software components, including third-party libraries. Conduct periodic security audits and vulnerability scans to identify and address potential issues. Default.

3. **Credentials Not Changed (Vul No. 6):**

- **Category:** PROTECT
- **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
- **Mitigation:** Align with PR. AA-01 (Identities and credentials for authorized users, services, and hardware are managed by the organization).

Enforce strong password policies and ensure that default credentials are changed during initial setup. Use multi-factor authentication (MFA) to enhance access security.

4. **Bugs and Code Performance Issues (Vul No. 10):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-06 (Secure software development practices are integrated, and their performance is monitored throughout the software development life cycle).

Establish a rigorous code review and testing process. Implement automated testing tools to detect bugs and performance issues early in the development cycle. Conduct regular performance evaluations and stress testing.

5. **Weak Access Controls (Vul No. 13):**

- **Category:** PROTECT
- **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
- **Mitigation:** Align with PR. AA-05 (Access permissions, entitlements, and authorizations are defined in a policy, managed, enforced, and reviewed, and incorporate the principles of least privilege and separation of duties).

Implement granular access controls and enforce the principle of least privilege. Regularly review and update access permissions. Use MFA and strong authentication mechanisms for sensitive systems.

6. **Weak Password Policies (Vul No. 17):**

- **Category:** PROTECT
- **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
- **Mitigation:** Align with PR. AA-01 (Identities and credentials for authorized users, services, and hardware are managed by the organization).

Develop and enforce strong password policies, including complexity requirements and regular password changes. Implement MFA for critical systems and accounts.

7. **Malicious Data in Training Datasets (Vul No. 21):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Implement data validation and cleansing procedures to detect and remove corrupted or malicious data. Use secure data sources and regularly review training data for integrity.

8. **Reverse-Engineering of the Model (Vul No. 22):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-05 (Installation and execution of unauthorized software are prevented).

Employ model protection techniques, such as encryption and obfuscation, to safeguard intellectual property. Regularly review and update security measures around model deployment

9. **Small, Crafted Inputs Causing Harmful Predictions (Vul No. 23):**

- **Category:** DETECT
- **Subcategory:** Continuous Monitoring (DE.CM)
- **Mitigation:** Align with DE.CM-09 (Computing hardware and software, runtime environments, and their data are monitored to find potentially adverse events).

Implement robust input validation and sanitization processes. Use anomaly detection mechanisms to identify and mitigate malicious inputs.

10. **GenAI Used for Social Engineering Attacks (Vul No. 31):**

- **Category:** DETECT
- **Subcategory:** Adverse Event Analysis (DE.AE)
- **Mitigation:** Align with DE.AE-02 (Potentially adverse events are analyzed to better understand associated activities).

Monitor and restrict the use of GenAI for generating sensitive or potentially harmful content. Implement detection mechanisms to identify misuse of the model.

11. **Algorithmic Manipulation (Vul No. 33):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)

- **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Apply security controls to prevent unauthorized access and modifications to the GenAI algorithm. Conduct regular audits to detect and address potential manipulations.

12. **Unauthorized Access to GenAI Application (Vul No. 34):**

- **Category:** PROTECT
- **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
- **Mitigation:** Align with PR. AA-06 (Physical access to assets is managed, monitored, and enforced commensurate with risk).

Strengthen access controls and implement logging and monitoring to detect unauthorized access attempts. Use encryption and secure authentication mechanisms.

1.8.2 Moderate-Risk Vulnerabilities Mitigation

13. **Weak Authentication (Vul No. 1):**

- **Category:** PROTECT
- **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
- **Mitigation:** Align with PR. AA-02 (Identities are proofed and bound to credentials based on the context of interactions).

Strengthen authentication mechanisms by enforcing strong password policies and implementing MFA. Regularly review and update authentication methods.

14. **Poorly Secured API Endpoints (Vul No. 3):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Secure API endpoints with proper authentication and authorization mechanisms. Implement rate limiting and input validation to protect against injection attacks.

15. **Lack of Timely Updates and Patches (Vul No. 5):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-02 (Software is maintained, replaced, and removed commensurate with risk).
- Develop a patch management strategy to ensure timely application of security updates.

Monitor for new vulnerabilities and apply patches as needed (Huang et al., 2024).

16. **Lack of Encryption for Data Exposure (Vul No. 7):**

- **Category:** PROTECT
- **Subcategory:** Data Security (PR.DS)
- **Mitigation:** Align with PR.DS-01 (The confidentiality, integrity, and availability of data-at-rest are protected).

Implement strong encryption for data at rest and in transit. Regularly review encryption practices to ensure they meet current security standards.

17. **Insufficient Monitoring and Auditing (Vul No. 14):**

- **Category:** DETECT
- **Subcategory:** Continuous Monitoring (DE.CM)
- **Mitigation:** Align with DE.CM-01 (Networks and network services are monitored to find potentially adverse events).

Enhance monitoring and auditing processes to detect and respond to suspicious activities. Implement comprehensive logging and regular reviews of access and activity logs.

18. **Compromised Third-Party Components (Vul No. 15):**

- **Category:** GOVERN
- **Subcategory:** Cybersecurity Supply Chain Risk Management (GV.SC)
- **Mitigation:** Align with GV.SC-07 (The risks posed by a supplier, their products and services, and other third parties are understood, recorded, prioritized, assessed, responded to, and monitored over the course of the relationship).

Evaluate and vet third-party software and components for security. Regularly update and patch third-party components and use trusted sources.

19. **Insufficient Rate Limiting on APIs (Vul No. 16):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Implement rate limiting and API gateway controls to manage and restrict API traffic. Monitor API usage for unusual patterns (Huang et al., 2024).

20. **Improper Configuration of Cloud Services (Vul No. 18):**

- **Category:** PROTECT
- **Subcategory:** Platform Security (PR.PS)
- **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Regularly review and audit cloud service configurations to ensure compliance with best practices. Implement security controls for data protection and access management.

21. **Unpatched Zero-Day Vulnerabilities (Vul No. 19):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-02 (Cyber threat intelligence is received from information sharing forums and sources).

Stay informed about emerging vulnerabilities and apply updates as soon as patches become available. Use threat intelligence to anticipate and prepare for zero-day threats (Krishnamurthy 2023).

22. **Lack of Adherence to Regulations (Vul No. 20):**

- **Category:** GOVERN
- **Subcategory:** Organizational Context (GV.OC)
- **Mitigation:** Align with GV.OC-03 (Legal, regulatory, and contractual requirements regarding cybersecurity — including privacy and civil liberties obligations — are understood and managed).

Ensure compliance with data protection and cybersecurity regulations through regular audits and updates to policies and procedures. Provide training to employees on regulatory requirements.

23. **Inherited Biases in Training Data (Vul No. 26):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).
- Regularly review and audit training data for biases. Implement techniques to detect and mitigate bias in model training.

24. **Inadequate or Biased Training Data (Vul No. 28):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).
- Use diverse and representative datasets for training. Regularly review data quality and make necessary adjustments to improve model accuracy and fairness (Krishnamurthy 2023).

25. **Unintentional Plagiarism (Vul No. 29):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Implement measures to detect and avoid replication of protected content. Use plagiarism detection tools and adhere to copyright laws.

26. **Privacy Invasion Through Outputs (Vul No. 32):**

- **Category:** PROTECT
- **Subcategory:** Data Security (PR.DS)
- **Mitigation:** Align with PR.DS-01 (The confidentiality, integrity, and availability of data-at-rest are protected).

Implement data anonymization techniques and review outputs to ensure they do not expose sensitive information.

27. **Ethical Violations (Vul No. 35):**

- **Category:** GOVERN
- **Subcategory:** Organizational Context (GV. OC)
- **Mitigation:** Align with GV. OC-01 (The organizational mission is understood and informs cybersecurity risk management).

Establish ethical guidelines for GenAI usage and ensure compliance through regular reviews and audits. Implement mechanisms to detect and address unethical outputs.

1.8.3 Low-Risk Vulnerabilities Mitigation

28. **Single Point of Failure in Server Hardware (Vul No. 8):**

- **Category:** PROTECT
- **Subcategory:** Technology Infrastructure Resilience (PR.IR)
- **Mitigation:** Align with PR.IR-03 (Mechanisms are implemented to achieve resilience requirements in normal and adverse situations).

Implement redundancy and failover solutions to mitigate the impact of hardware failures. Regularly test disaster recovery plans.

29. **Insufficient Encryption on Wireless Communication (Vul No. 9):**

- **Category:** PROTECT
- **Subcategory:** Data Security (PR.DS)

- **Mitigation:** Align with PR.DS-02 (The confidentiality, integrity, and availability of data-in-transit are protected).
- Use strong encryption protocols for wireless communications. Regularly review and update encryption practices.

30. **Lack of Disaster Recovery Plans (Vul No. 12):**

- **Category:** RESPOND
- **Subcategory:** Incident Recovery Plan Execution (RC.RP)
- **Mitigation:** Align with RC.RP-01 (The recovery portion of the incident response plan is executed once initiated from the incident response process).

Develop and test comprehensive disaster recovery plans. Ensure plans address various scenarios and include regular updates.

31. **Performance Degradation Over Time (Vul No. 24):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-04 (Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded).
- Implement monitoring tools to track model performance and conduct regular evaluations. Update models as necessary to address performance issues.

32. **Generation of Nonsensical Outputs (Vul No. 25):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Continuously review and refine the model to improve output quality. Implement feedback mechanisms to detect and address nonsensical outputs.

33. **Difficulty in Interpreting Model Outputs (Vul No. 27):**

- **Category:** IDENTIFY
- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Enhance model explain ability and provide clear documentation on output interpretation. Use visualization tools to aid in understanding outputs.

34. **The Model Performs Poorly on New Data (Vul No. 30):**

- **Category:** IDENTIFY

- **Subcategory:** Risk Assessment (ID.RA)
- **Mitigation:** Align with ID.RA-04 (Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded).

Implement continuous learning and adaptation strategies to improve model performance on new data. Regularly update the training data to reflect changing patterns.

References

Asaad, R.R. and Saeed, V.A., 2022. A Cyber Security Threats, Vulnerability, Challenges and Proposed Solution. *Applied computing Journal*, pp.227-244.

Blank, R. and Gallagher, P., 2012. Nist special publication 800-30 revision 1 guide for conducting risk assessments. *National Institute of Standards and Technology*.

Dimakopoulou, A. and Rantos, K., 2024. Comprehensive Analysis of Maritime Cybersecurity Landscape Based on the NIST CSF v2. 0. *Journal of Marine Science and Engineering*, *12*(6), p.919.

Huang, K., Yeoh, J., Wright, S. and Wang, H., 2024. Build your security program for genai. In *Generative AI Security: Theories and Practices* (pp. 99-132). Cham: Springer Nature Switzerland.

Katsumata, P., Hemenway, J. and Gavins, W., 2010, October. Cybersecurity risk management. In *2010-MILCOM 2010 Military Communications Conference* (pp. 890-895). IEEE.

Möller, D.P., 2023. NIST cybersecurity framework and MITRE cybersecurity criteria. In *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices* (pp. 231-271). Cham: Springer Nature Switzerland.

Safitri, E.H.N. and Kabetta, H., 2023, August. Cyber-Risk Management Planning Using NIST CSF V1. 1, ISO/IEC 27005: 2018, and NIST SP 800-53 Revision 5 (A Study Case to ABC Organization). In *2023 IEEE International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs)* (pp. 332-338). IEEE.

Central Queensland university (CQU) Intellectual property

COIT20265: NETWORK AND INFORMATION SECURITY PROJECT (HT2, 2024)

FINAL REPORT

PROJECT:

GENERATIVE AI: NAVIGATING SHORT-TERM SKEPTISM AND LONG-TERM PROMISE

UNDER THE ESTEEMED GUIDANCE OF

UNIT CO-ORDINATOR: FARIZA SABRINA

MENTOR: DR. AHMEDI AZRA

TEAM MEMBERS:

BASANTA ADHIKARI (12211752)

BHUWAN THAPA (12196590)

KIRAN BHUSAL (12211570)

PRATIK SHINGH DHAMI (12209929)

Table of Contents

1. System Model Diagram

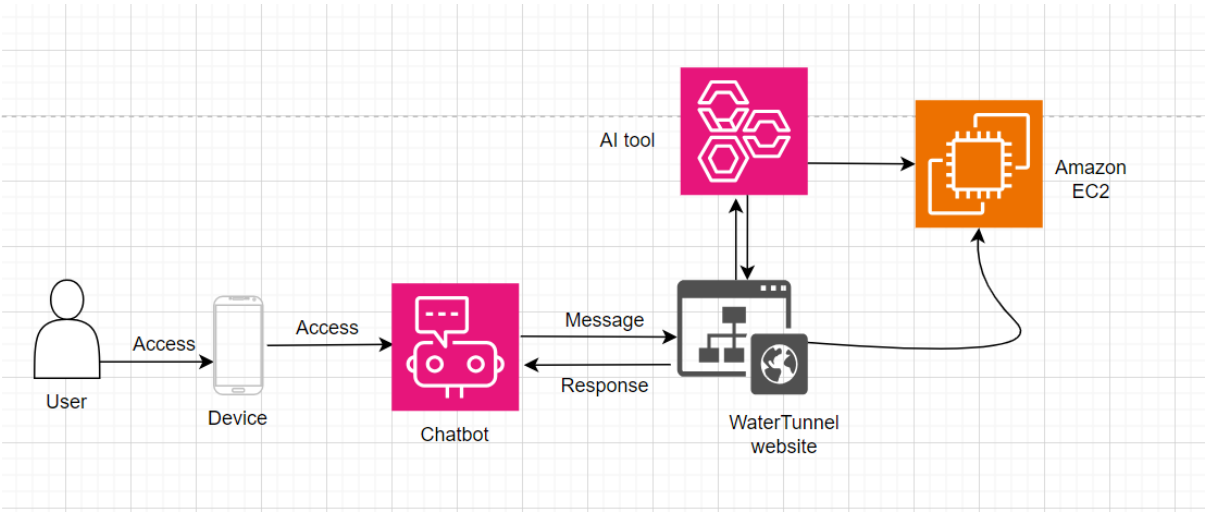


Figure-1: System Design

In the above figure, AI is associated to the website detects the threats and vulnerabilities in the cybersecurity. The website and AI tool both are hosted on EC2 instances under AWS webserver. In addition, while implementing chatbot a user sends an enquiry message through a website, the chatbot will respond to the message from customers. The aim main of the implementing AI tool in this project is to detect cybersecurity vulnerabilities. The AI tool is designed to detect vulnerabilities of the website. The user or security leasers input the URL of the website, AI tool generate vulnerability report that assists and alerts security leaders to make safe secure the company systems. This report will discuss the security proposals for the entire company system just below. All these systems and data is stored and monitored by AWS cloud services.

2. Design of Network / Security Architecture

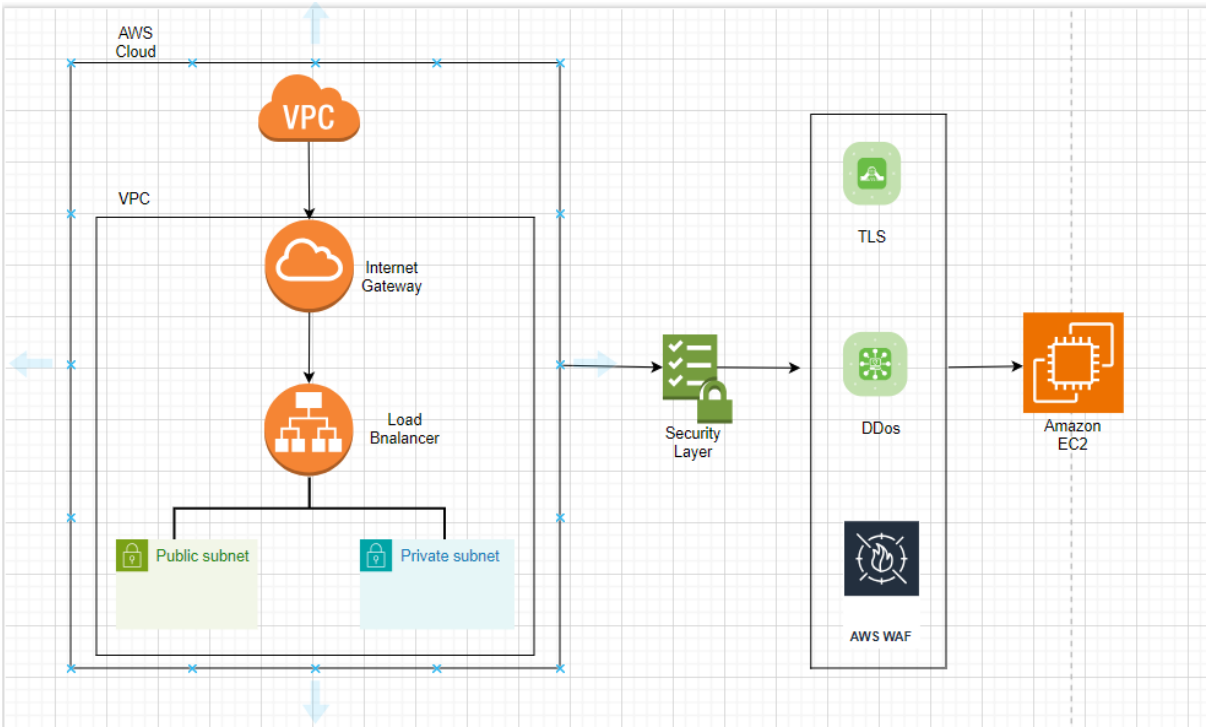


Figure-2: Network/security design

Network Architecture/Protocols/Algorithms

While deploying AI tool and websites on AWS cloud servers, the network architecture plays a crucial role for ensuring secure and efficient communication between clients and servers. Here is a description of the key components of the network architecture:

1. AWS Cloud: The entire system is deployed on the AWS cloud infrastructure, which provides scalability, reliability, and a wide range of services to support the deployment of applications and services.

2. Virtual Private Cloud (VPC): A VPC is created to logically isolate the network resources of the system within the AWS cloud. It allows to define a virtual network environment with its IP address range, subnets, route tables, and network gateways (Patibandla, K.R., 2024).

3. Internet Gateway: An Internet Gateway is attached to the VPC to enable communication between instances in the VPC and the internet. It allows inbound and outbound traffic to and from the internet, facilitating the user's interaction on the website.

4. Load Balancer: A load balancer is used to distribute incoming traffic across multiple AWS web servers hosting the website and AI tool. This helps in load distribution, improves availability, and provides fault tolerance by ensuring that no single server is overwhelmed with traffic.

5. Public Key and Private Key:

- Key Infrastructure (PKI) is implemented in the network architecture to secure communications and authenticate entities. “Public keys are used for encryption and verification, while private keys are used for decryption and signing” (Patibandla, K.R., 2024).
- Public and private key pairs are used for establishing secure connections, such as SSL/TLS connections, between clients and servers. The public key is shared openly, while the private key is kept secure and known only to the server.
- Public and private keys play a crucial role in ensuring secure communication channels and protecting sensitive data transmitted over the network.

Security Architecture/Protocols/Algorithms

In the deployment of websites and AI tool hosted on AWS cloud servers, a comprehensive security architecture is essential to protect the system from various threats and vulnerabilities. Here is a description of the key components of the security architecture:

1. Transport Layer Security (TLS):

    - TLS is implemented to secure communication between clients (e.g. website visitors) and servers (AWS web servers hosting the website).
    - “TLS encrypts data in transit, ensuring that sensitive information exchanged between clients and servers is protected from eavesdropping and tampering” McKay, K. and Cooper, D. (2017, p. 01).
    - By using TLS, the security architecture ensures the confidentiality and integrity of data transmitted over the network.

2. Distributed Denial of Service (DDoS) Protection:

    - DDoS protection mechanisms are implemented to defend against DDoS attacks that aim to disrupt the availability of the customers interaction to the website.

- Utilizing AWS Shield, a managed DDoS protection service, helps safeguard the system from volumetric and application layer DDoS attacks by detecting and mitigating malicious traffic.

3. AWS Web Application Firewall (WAF):

    - AWS WAF is deployed to protect the web application (website and AI tool) from common web exploits and vulnerabilities (Lakhno et al., 2022).
    - It allows security rules to be defined to filter and monitor incoming web traffic, blocking malicious requests before they reach the web servers.
    - AWS WAF works in conjunction with the load balancer to provide an additional layer of defence against SQL injection, cross-site scripting (XSS), and other security threats.

with all those above-mentioned network and security architecture, we can enhance security of the website and AI tools. Even tools will be able to notice unethical cyber threats analysing the intension of the users. Including above mentioned security protocols and algorithms and implementing network access control lists within the VPC we can inbound and outbound traffic at the traffic at the subnet level. Configuring ACL we can restrict traffic based on IP addresses and ranges that will help to effectively restrict access from unauthorised access.

3. Website design and development

Regarding the technical progress of the project till now, we have developed and designed a website for Automotive industry specific to WaterTunnel car wash company. For developing websites, deployed Wordpress certified by Bitnami by using AWS server involving several steps. The main reasons for choosing aws cloud service for deploying wordpress are security and cost-effectiveness. When it completed the process of launching instances in EC2 we received a public IPv4 address and login username and password for the wordpress website.

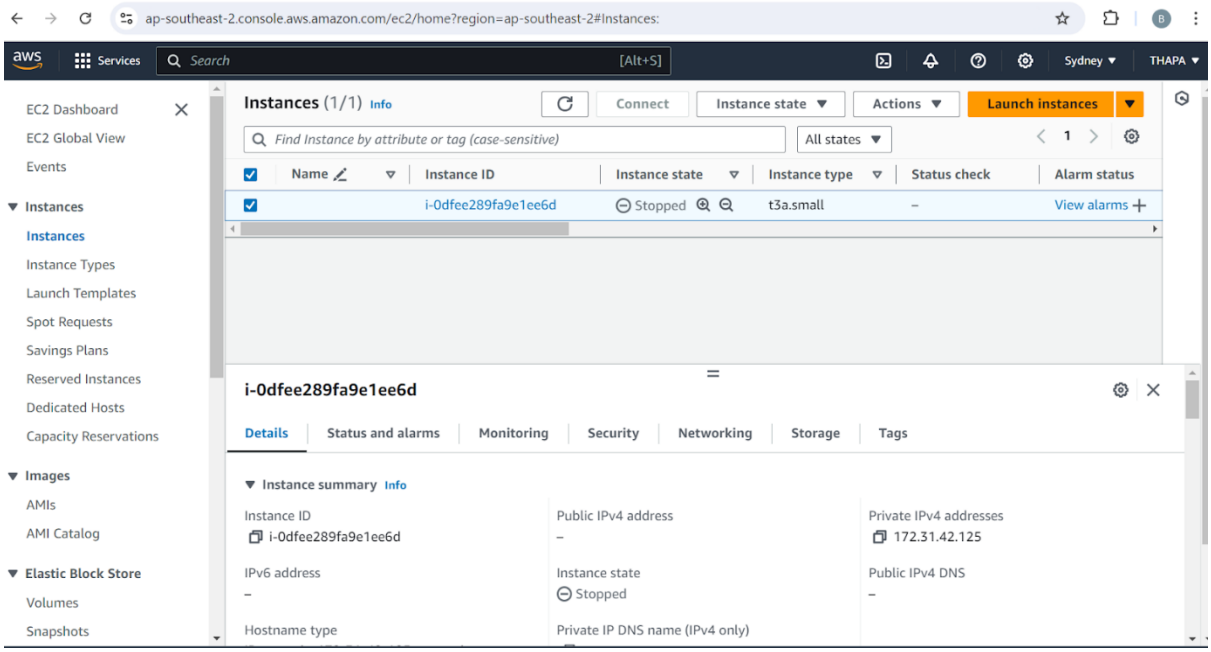


Fig 3: Launching Bitnami WordPress in AWS.

When creating websites then collected different resources from online sources and some of them are self-designed. The website contains everything that needs to be a perfect website. The website is all about company's information and car wash features. The website includes service details, Wash Menu, contact details etc. Here are some glimpses of websites.

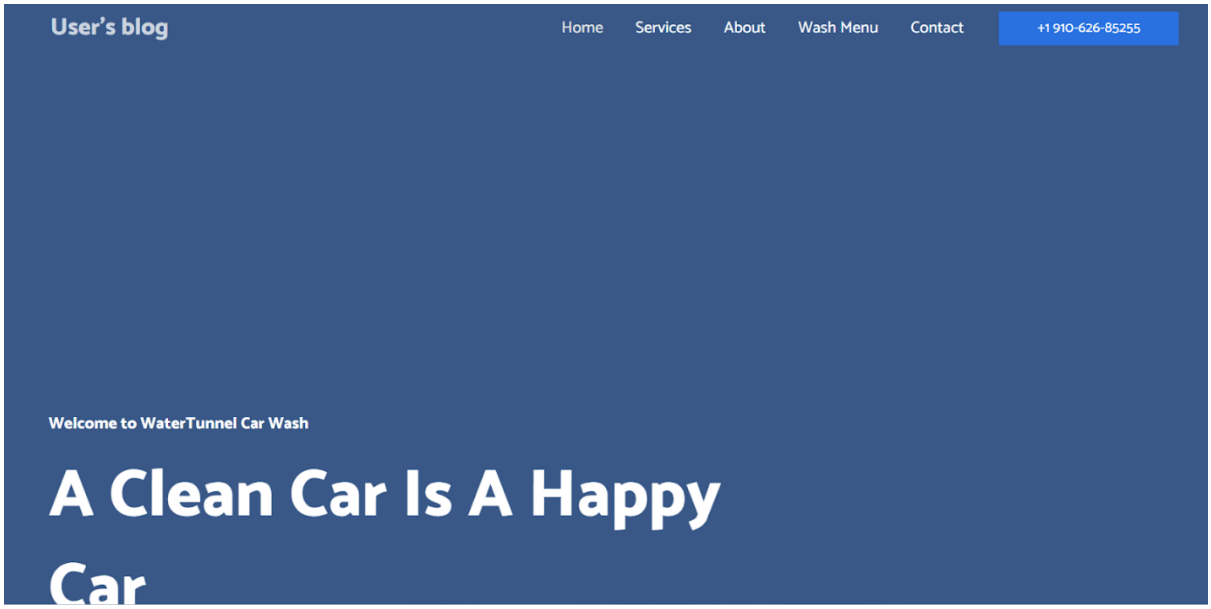


Fig 4: Company Website

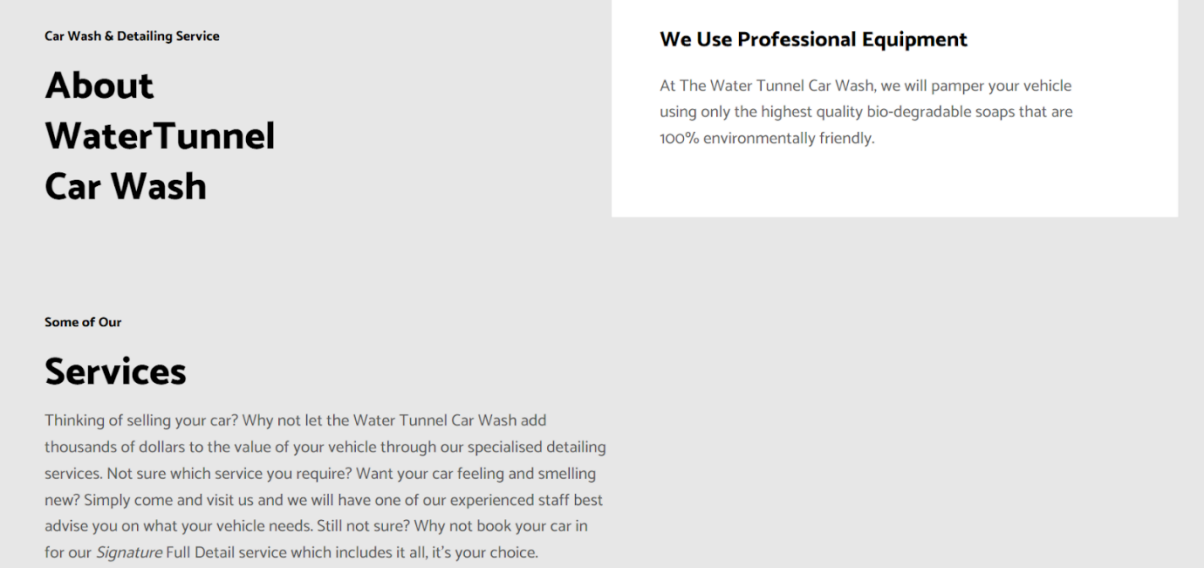


Fig 5: Company Website

3.1 Deployment process of WordPress website

The developed website is deployed into EC2 instance under AWS server for scalability, reliability, security and cost-effectiveness etc.

Deployment process of website in EC2 instance:

1. Login to AWS server.
2. Select “EC2” instances from all services.
3. Select “lunch instance” under EC2.
4. Select “browse more AMIs” by default.
5. Select “AWS Marketplace AMIs” under AMI.
6. Select desired WordPress website.
7. Name website “WordPress or any other name” under key pair.
8. Select “lunch instances” at the bottom and wait for couple of minutes
9. Again, go back to instances under EC2, you will see running instances
10. Select “running instances” you will see details of the website such as public IP and private IP.
11. Copy “public IP” and open in new tab you will see website format and design it.
12. For login username and password for WordPress, select running “instance ID”.
13. Select “action” and again select “monitor and troubleshoot”.
14. Seach for user Id and password to login in WordPress site.

4. AI Tool to Detect Vulnerabilities in Cybersecurity

Gen AI plays an important role in providing security layers for the website or any other application. As this report has already discussed about the problems, solutions, benefits and limitations of GenAI while detecting the vulnerabilities in cybersecurity. Here, the aim is to generate such AI tool which detects Vulnerabilities on the website that we already designed.

The AI will be based on the following algorithms:

1. Data collection:
    - Gather website content including HTML, CSS, JavaScript and backend code for analysis.
    - Extract security related metadata such as headers, cookies, and response codes.
2. Input processing:
    - Tokenize and purse the website code into analysable units.
    - Normalise data to eliminate any inconsistencies.
3. Vulnerability detection:
    - Pattern matching: Identify known vulnerability pattern using predefined rule sets (e.g SQL injection, XSS, CSRF).
    - Anomaly detection: Use AI/ML models to detect unusual patterns or deviations in the website's structure and that may indicate new vulnerabilities.
4. Security Compliance Check:
    - Compare website configurations and code against industry security standards (e.g., OWASP Top 10) to ensure compliance.
    - Generate a compliance report highlighting areas of concern.
5. Recommendation Generation:
    - Based on detected vulnerabilities, generate actionable security recommendations.
    - Prioritize vulnerabilities by severity and impact.
6. Reporting:
    - Generate a detailed security report with findings, risk assessments, and recommended actions.
    - Provide options for automatic or manual fixes based on severity.

7. Deployment:
    - Integrate the Al tool into a continuous integration/continuous deployment (CI/CD) pipeline.
    - Deploy the Al tool on AWS to enable scalability and access control.
8. Feedback Loop:
    - Continuously update the Al models with new vulnerability data and improve detection algorithms based on user feedback.

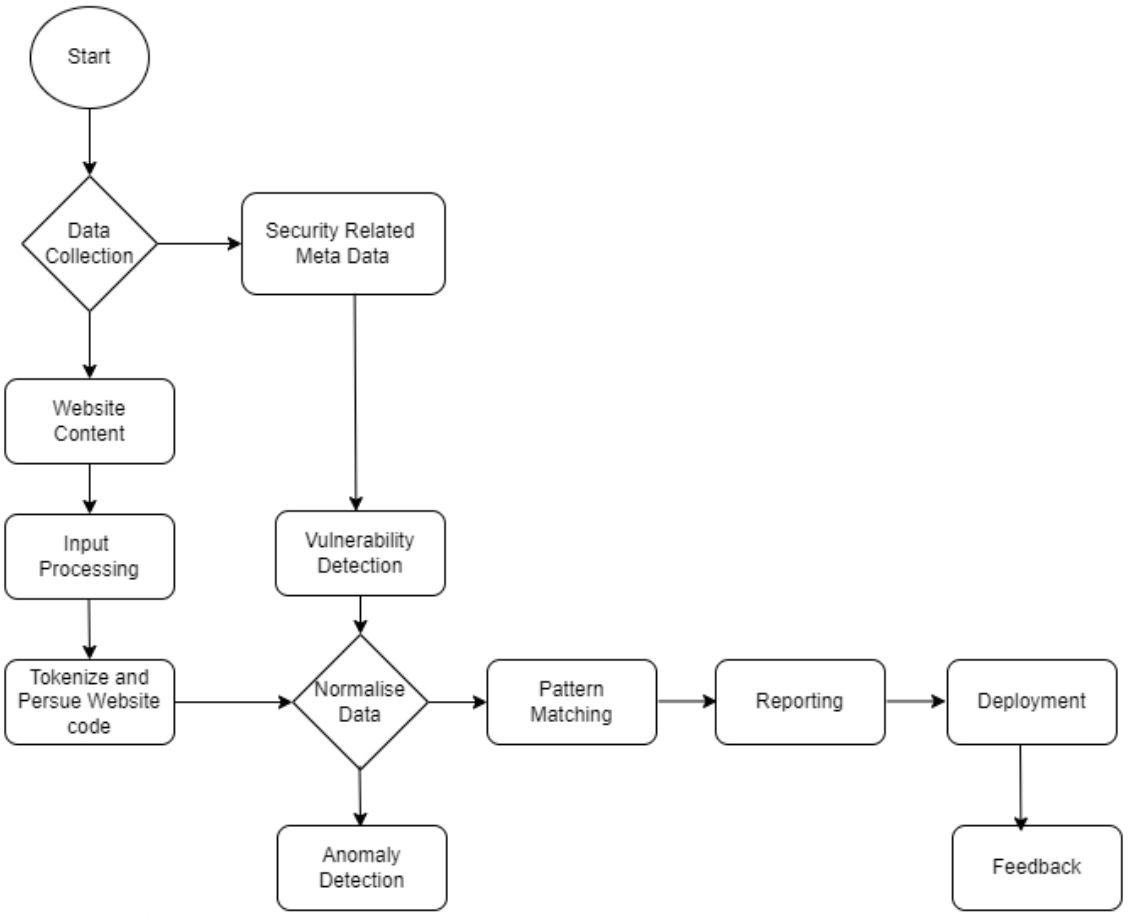


Fig 6: Flowchart for GenAI Tool

4.1 Software Requirements Specifications (SRS)

Here this topic defines the functional and non-functional requirements for the web application vulnerability scanner. The system is developed to scan website for common vulnerabilities like SQL injection, cross-site scripting (XSS), SSRF, RCE and many more. The scanner will help developer and security leaders identify and assess vulnerabilities in their website.

The web application vulnerability will allow user to enter a URL and scan various security vulnerabilities of the website. The system checks foe vulnerabilities across common endpoints, Analyse server responses and classify the results. The vulnerabilities logged and reported in the comprehensive format. The scanner also performs brute force login attempts and anomaly detection in server responses.

SQL Injection: A code injection technique that allow unauthorized access to a database, enable attackers to view modify and delete data.
XSS (Cross Site Scripting): When web application accepts user input and includes in webpages without proper validation or sanitization the malicious script is permanently stored in webpages.
Server Site Request Forgery (SSRF): When a user clicks on vulnerable URL, that allows attackers to access the information of the users.
Cross Site Request Forgery (CSRF): An attack where an attackers trapped users to perform actions on web application by sending them links.
Brute force Attack: an attack that used to gain unauthorised access to the system by systematically trying every possible combination of password.

Functional Requirements

1. The system must accept URL as input from user.
2. The system will check common paths such as s (e.g., /wp-admin, /wp-login.php, /phpinfo.php) for vulnerabilities.
3. The system will classify the type of vulnerability using a pre-trained machine learning model such as (e.g., SQL Injection, XSS, CSRF, etc.).
4. The system will attempt brute force login with common username and password combinations.
5. The system will detect anomalies in the server responses based on a trained anomaly detection model.
6. The system will log vulnerabilities and errors to an HTML file.

Interface requirements

1. User interface: valid URL for scanning and HTML pages showing vulnerabilities results, classification and report.
2. No hardware requirements.
3. Software requirements: python, flask that serves as a web framework, TensorFlow used for machine learning classification, Hugging Face Transformers for pre-trained model integration.

4.2 Deployment of AI tool onto AWS server

The developed AI tool is deployed into EC2 instance under AWS server for scalability, reliability, security and cost-effectiveness etc.

Deployment process of AI tool

1. Prepare AI model
- Choose framework e.g TenssorFlow

2. Create deployment package

- Setup virtual environment

Type: python -m venv myenv

source myenv/bin/activate

- Install necessary libraries

Type: pip install tensorflow

pip install numpy

- Package Your Code: Create folder and copy AI model and code

Type: mkdir my-deployment-package

cp -r mymodel.h5 my-deployment-package/

cp lambda_function.py my-deployment-package/

cd my-deployment-package

zip -r ../my-deployment-package.zip .

cd ..

3. Login to AWS management console

4. Navigate to lambda

- Click on "Create function."

- Choose "Author from scratch."

- Fill in the function name, select the runtime (e.g., Python 3.x), and set execution permissions (IAM role).

5. Upload Your Deployment Package

- In the "Function code" section, select "Upload a .zip file" and upload my-deployment-package.zip.

6. Set necessary environment variables in lambda configuration.

7. Increase memory and timeout if necessary.

8. Go to “Test Tab” and create a new test event.

9. Run the test and check the output.

Steps for Running AI tool

1. Login to AWS server.
2. Access the EC2 instance.
3. Navigate to AI Scanner:
- Select the EC2 instance that host AI Scanner.
4. Start the instance:
- If the instance is in a stopped state, go to the instance state and click start.

5. Connect the EC2 instance:
- Once the instance is running, select the instance and click connect.
6. Establish a connection:
- Choose the method to connect to your instance (EC2 instance connect or other) and click connect at the bottom to initiate the session.
7. Navigate to Web Scanner Directory:
- In the terminal type the following command to navigate to the Web Scanner directory: cd Web_Scanner/
8. To see deployed python code type: cat app.py
9. Activate the Virtual Environment by typing: source venv/bin/activate
10. Start the Web Server:
- Start the AI Scanner application using Gunicorn with following command: gunicorn -w 4 -b 0.0.0.0:5000 -timeout 1200 app:app

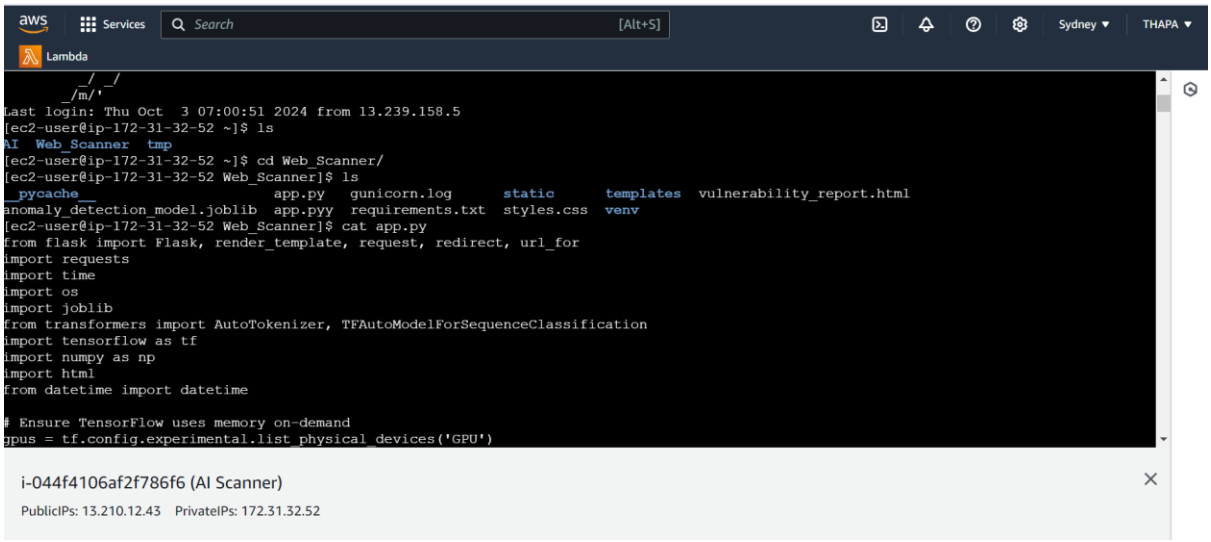


Fig 7: Running AI Scanner in Virtual Environment

4.3 AI to Identify Cybersecurity Vulnerabilities

This project has completed the basic task of developing AI which is able to detect cybersecurity vulnerabilities, and checked using URL of the website and got positive result. The AI is basically generated in VS code using python programming language and used the basic models available on python library and used AI ML model such as isolation forest and tensor flow for anomaly detection and also used pre trained AI model Microsoft/codebert-base for making our task easy. Basically, the generated AI isa able to detect vulnerabilities on the different areas of the websites through SQL injection, Cross site scripting, brute-force attack and we have tried Cross-site Request Forgery (CSRF) and Server-site Request Forgery (SSRF). The result has been shown below.

Vulnerability Scan Report for http://52.63.41.222/

Path	Description	Severity Level
wp-admin/install.php	[VULNERABLE] http://52.63.41.222/wp-admin/install.php - WordPress installation script is accessible. (Type: XSS, Confidence: 0.55)	
wp-admin/install.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-admin/install.php with payload: http://localhost [ANOMALOUS RESPONSE DETECTED]	
wp-admin/install.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-admin/install.php with payload: http://127.0.0.1	
wp-admin/install.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-admin/install.php with payload: http://169.254.169.254/latest/meta-data/ [ANOMALOUS RESPONSE DETECTED]	
wp-login.php	[VULNERABLE] http://52.63.41.222/wp-login.php - WordPress login page is exposed. (Type: XSS, Confidence: 0.56)	
wp-login.php	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-login.php with payload: <script>alert('XSS')</script>	

Fig 8: Vulnerability Scan Report

wp-login.php	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-login.php with payload:
wp-login.php	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-login.php with payload: <div onmouseover='alert("XSS")'>Hover</div>
wp-login.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-login.php with payload: http://localhost
wp-login.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-login.php with payload: http://127.0.0.1
wp-login.php	[VULNERABLE] SSRF possible at: http://52.63.41.222/wp-login.php with payload: http://169.254.169.254/latest/meta-data/
wp-login.php	[VULNERABLE] RCE possible at: http://52.63.41.222/wp-login.php with payload: phpinfo()
wp-login.php	[VULNERABLE] RCE possible at: http://52.63.41.222/wp-login.php with payload: system("ls")
wp-login.php	[VULNERABLE] RCE possible at: http://52.63.41.222/wp-login.php with payload: system("cat /etc/passwd")
wp-content/debug.log	[VULNERABLE] XSS possible at: http://52.63.41.222/wp-content/debug.log with payload:

Fig 9: Vulnerability Scan Report

5. List of Issues and Challenges and Mitigation

Issues and Challenges	**Mitigation**
Finding official WordPress.	For finding official WordPress, went through most of AMI in AWS most of them were paid, chose WordPress certified by Bitnami because of its cost-effectiveness.
Selecting security group configuration while launching.	It confused about whether to select SSH, HTTP or HTTPS. Later, did google research and found out to select all three.
Finding WordPress admin password.	We waited up to 1 hr after completing launch in EC2 to receive username and password.
Unexpected cost.	The cost for launching WordPress will be deducted per hour rate so we must visit AWS web server regular.

While creating the Gantt chart, errors indicated by red lines appeared on the start and end dates of some tasks	We resolved this issue by adjusting the predecessors and re-entering the data, which corrected the errors and allowed us to generate the Gantt chart successfully.
It was difficult to select an appropriate industry and identify the relevant stakeholders.	We chose the automotive industry, specifically a water tunnel car wash. With the help of the manager, we gathered information about the stakeholders.
Scanning urls or multiple tests cases network latency can cause timeouts.	Use reasonable timeout values for 'requests.get' and 'requests.post' that we have used.
Automated vulnerability scanners can produce false positives.	Refined payloads for SQL injection and XSS have been used.
Using an anomaly detection model can yield false positives.	Training data for the anomaly detection model is representative of both normal and abnormal responses.
Training data for the anomaly detection model is representative of both normal and abnormal responses.	Our program includes memory-efficient data structures.
Pre-trained CodeBERT model might not always produce accurate result.	Fine-tune the CodeBERT model with a dataset of labelled vulnerability reports specific to our use case.

References

Patibandla, K.R., 2024. Design and Create VPC in AWS. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, *1*(1), pp.273-282.

McKay, K. and Cooper, D., 2017. *Guidelines for the selection, configuration, and use of transport layer security (TLS) implementations* (No. NIST Special Publication (SP) 800-52 Rev. 2 (Draft)). National Institute of Standards and Technology.

Lakhno, V., Blozva, A., Kasatkin, D., Chubaievskyi, V., Shestak, Y., Tyshchenko, D. and Brzhanov, R., 2022. Experimental studies of the features of using waf to protect internal services in the zero trust structure. *J Theor Appl Inf Technol*, *100*(3), pp.705-721.