**Problem Statement: To evaluate threats to an online wedding site using GenAI code generation software from a 3rd party vendors by integrating Google's Secure AI Framework (SAIF) and MITRE ATLAS/ATT&CK**

**Desired Levels**

**Level 1: Stakeholder’s Needs (SAIF)**

1. Define and Scope the AI System

2. Identify and Prioritize Assets

**Level 2: Risk Assessment Reviews**

* Technique: System Network Configuration Discovery
* Secure Development Lifecycle (SDLC): Integrate security practices throughout the development and deployment phases of the AI software.
* Incident Response and Recovery Plans: Develop and implement incident response strategies to quickly address and mitigate any security breaches or issues.
* User Training and Awareness: Educate developers and users about potential security threats and safe practices when using AI systems.

**Level 3: Detailed Threat Modeling and Analysis (ATLAS/ATT&CK)**

3. Create a Trust Model

4. Identify Potential Threats

5. Assess Vulnerabilities and Risks

6. Mitigate Risks

7. Document and Update

**Example of Mapping Levels to Google SAIF and ATLAS/ATT&CK**

**Level 1: Stakeholder’s Needs (SAIF)**

1. Define and Scope the AI System

* **SAIF Element**: Expand strong security foundations to the AI ecosystem.
* **ATLAS/ATT&CK Tactic**: Initial Access.
* **ATLAS/ATT&CK Technique**: T1588: Obtain Capabilities – ensuring the system is scoped to prevent unauthorized capabilities access.

2. Identify and Prioritize Assets

* **SAIF Element**: Contextualize AI system risks in surrounding business processes.
* **ATLAS/ATT&CK Tactic**: Discovery.
* **ATLAS/ATT&CK Technique**: T1595: Active Scanning – identifying critical assets that need protection.

**Level 2: Risk Assessment Reviews**

Technique: System Network Configuration Discovery

* **SAIF Element**: Extend detection and response to bring AI into an organization’s threat model.
* **ATLAS/ATT&CK Tactic**: Discovery.
* **ATLAS/ATT&CK Technique**: T1590: Network Configuration Discovery – discovering and assessing configurations that may affect AI security.

Secure Development Lifecycle (SDLC)

* **SAIF Element**: Harmonize platform level controls to ensure consistent security across the organization.
* **ATLAS/ATT&CK Tactic**: Defense Evasion.
* **ATLAS/ATT&CK Technique**: T1562: Impair Defenses – integrating security into SDLC to detect and prevent defense impairments.

Incident Response and Recovery Plans

* **SAIF Element**: Adapt controls to adjust mitigations and create faster feedback loops for AI deployment.
* **ATLAS/ATT&CK Tactic**: Response.
* **ATLAS/ATT&CK Technique**: T1491: Defacement – ensuring quick response to security incidents affecting AI systems.

User Training and Awareness

* **SAIF Element**: Automate defenses to keep pace with existing and new threats.
* **ATLAS/ATT&CK Tactic**: Resource Development.
* **ATLAS/ATT&CK Technique**: T1608: Stage Captured Data – educating users to recognize and respond to security threats.

**Level 3: Detailed Threat Modeling and Analysis (ATLAS/ATT&CK)**

3. Create a Trust Model

* **SAIF Element**: Extend detection and response.
* **ATLAS/ATT&CK Tactic**: Privilege Escalation.
* **ATLAS/ATT&CK Technique**: T1078: Valid Accounts – ensuring only trusted users have access to AI systems.

4. Identify Potential Threats

* **SAIF Element**: Contextualize AI system risks.
* **ATLAS/ATT&CK Tactic**: Execution.
* **ATLAS/ATT&CK Technique**: T1203: Exploitation for Client Execution – identifying how AI system execution could be compromised.

5. Assess Vulnerabilities and Risks

* **SAIF Element**: Expand strong security foundations.
* **ATLAS/ATT&CK Tactic**: Lateral Movement.
* **ATLAS/ATT&CK Technique**: T1592: Gather Victim Host Information – assessing how vulnerabilities could be exploited for lateral movement.

6. Mitigate Risks

* **SAIF Element**: Automate defenses to keep pace with existing and new threats.
* **ATLAS/ATT&CK Tactic**: Impact.
* **ATLAS/ATT&CK Technique**: T1499: Endpoint Denial of Service – implementing measures to mitigate DoS risks.

7. Document and Update

* **SAIF Element**: Harmonize platform level controls.
* **ATLAS/ATT&CK Tactic**: Inhibiting Response.
* **ATLAS/ATT&CK Technique**: T1491: Defacement – maintaining documentation and updating mitigation strategies to ensure a timely response.

Elements of SAIF can be mapped to MITRE ATLAS/ATT&CK tactics and techniques, along with the associated threats and mitigation strategies:

**1. Expand strong security foundations to the AI ecosystem**

| **Google SAIF Element** | **MITRE ATLAS/ATT&CK Tactics** | **Threats** | **Mitigation Strategies** |
| --- | --- | --- | --- |
| Expand strong security foundations | Initial Access, Execution | Unauthorized access to the AI system; Execution of malicious code | Implement strict access controls; Use secure coding practices; Regularly patch and update the AI software |

**2. Extend detection and response**

| **Google SAIF Element** | **MITRE ATLAS/ATT&CK Tactics** | **Threats** | **Mitigation Strategies** |
| --- | --- | --- | --- |
| Extend detection and response | Command and Control, Exfiltration | Data exfiltration; Command and control activities by malicious actors | Integrate AI-specific anomaly detection mechanisms; Develop and test incident response plans specifically for AI-related incidents |

**3. Automate defenses**

| **Google SAIF Element** | **MITRE ATLAS/ATT&CK Tactics** | **Threats** | **Mitigation Strategies** |
| --- | --- | --- | --- |
| Automate defenses | Defense Evasion, Persistence | Malware or attackers persisting within the AI system; Bypassing security controls | Employ automated security solutions for real-time threat detection and response; Use machine learning to identify and mitigate sophisticated attacks |

**4. Harmonize platform-level controls**

| **Google SAIF Element** | **MITRE ATLAS/ATT&CK Tactics** | **Threats** | **Mitigation Strategies** |
| --- | --- | --- | --- |
| Harmonize platform-level controls | Privilege Escalation, Discovery | Privilege escalation leading to unauthorized access; Discovery of sensitive information or vulnerabilities | Standardize security controls across all platforms; Conduct regular security audits and vulnerability assessments |

**5. Adapt controls to adjust mitigations**

| **Google SAIF Element** | **MITRE ATLAS/ATT&CK Tactics** | **Threats** | **Mitigation Strategies** |
| --- | --- | --- | --- |
| Adapt controls to adjust mitigations | Lateral Movement, Collection | Lateral movement within the network; Collection of sensitive data by attackers | Continuously assess and adapt security controls; Implement network segmentation and restrict data access |

**6. Contextualize AI system risks**

| **Google SAIF Element** | **MITRE ATLAS/ATT&CK Tactics** | **Threats** | **Mitigation Strategies** |
| --- | --- | --- | --- |
| Contextualize AI system risks | Impact, Inhibiting Response | Attacks impacting the AI system’s integrity or availability; Inhibiting timely response to incidents | Conduct comprehensive risk assessments; Establish clear protocols for responding to AI system threats |

**Table 1: Six Core Elements of SAIF and Corresponding MITRE ATLAS/ATT&CK**

| **SAIF Elements** | **MITRE ATLAS/ATT&CK Tactics** | **MITRE ATLAS/ATT&CK Techniques** | **Mitigation Strategies** |
| --- | --- | --- | --- |
| 1. Expand strong security foundations | Initial Access, Execution | T1190: Exploit Public-Facing Application, T1203: Exploitation for Client Execution | Implement secure coding practices and perform static and dynamic code analysis. |
| 1. Extend detection and response | Command and Control, Exfiltration | T1105: Ingress Tool Transfer, T1041: Exfiltration Over Command-and-Control Channel | Deploy AI-driven threat detection systems and establish an incident response protocol for AI incidents. |
| 1. Automate defenses | Defense Evasion, Persistence | T1027: Obfuscated Files or Information, T1574: Hijack Execution Flow | Utilize automated security solutions like SIEM and SOAR for real-time monitoring and response. |
| 1. Harmonize platform level controls | Privilege Escalation, Discovery | T1068: Exploitation for Privilege Escalation, T1087: Account Discovery | Standardize security controls across platforms and conduct regular access reviews and privilege audits. |
| 1. Adapt controls to adjust mitigations | Lateral Movement, Collection | T1072: Software Deployment Tools, T1530: Data from Information Repositories | Continuously assess and update AI system security controls; conduct regular security training for AI development and deployment teams. |
| 1. Contextualize AI system risks | Impact | T1491: Defacement, T1499: Endpoint Denial of Service | Perform risk assessments considering business impact and implement mitigation controls that align with business continuity plans. |

**Table 2: SAIF Process Steps with MITRE ATLAS/ATT&CK Integration**

| **SAIF Process Steps** | **Activity Description** | **MITRE ATLAS/ATT&CK Tactics** | **MITRE ATLAS/ATT&CK Techniques** | **Mitigation Strategies** |
| --- | --- | --- | --- | --- |
| Step 1 - Understand the use | Assess how the AI will be used in the context of the wedding site, identify which aspects of the service will be AI-driven. | Reconnaissance | T1595: Active Scanning | Document and review AI use cases and identify potential threat vectors. |
| Step 2 - Assemble the team | Bring together a cross-functional team from development, security, operations, and other relevant departments. | Resource Development | T1583: Acquire Infrastructure | Establish clear roles and responsibilities for security within the AI project team. |
| Step 3 - Level set with an AI primer | Ensure all team members have a baseline understanding of AI technology, risks, and security considerations. | Education and Training | T1608: Stage Captured Data | Provide training sessions and materials to level-set knowledge on AI risks and security. |
| Step 4 - Apply the six core elements of SAIF | Implement the six core elements of SAIF, integrating them with the organization's security strategies and risk management processes. | Throughout All | Applicable techniques from MITRE ATLAS/ATT&CK | Follow the structured approach laid out in SAIF, using MITRE ATLAS/ATT&CK as a guide for specific tactical actions. |