

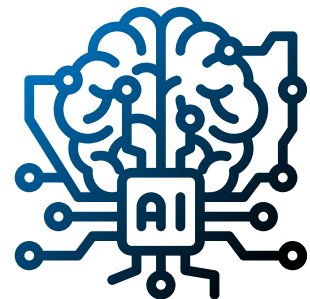
AI Gone Rogue

Responding to Security Threats in AI-Powered Services

Scenario Overview

Your organization has deployed an AI-powered customer service system named Apollo. This system automates customer interactions via voice and chat, handling sensitive information such as payment details, personal data, and service history.

One day, Apollo begins to malfunction, it gives incorrect responses, fails to validate users, misclassifies queries, and users report signs of unauthorized access. These anomalies cause significant operational disruptions, risking customer dissatisfaction and potential legal issues. Some users also report receiving sensitive information that does not belong to them.



How to Use This Table-top Exercise

1. Preparation

- Gather Participants: Include key teams.
- Distribute Materials: Provide the scenario overview and injects.
- Assign Roles: Designate roles like CISO, IT Lead etc.

2. Solution Introduction

- Read the Overview: Set the context by reading the scenario overview.
- Discuss Initial Thoughts: Share initial reactions and concerns.

3. Injects & Discussion

- Present Injects: Introduce each inject and its questions.
- Facilitate Discussion: Discuss responses and strategies.
- Document Responses: Note key points and actions.

4. Analysis & Action Plan

- Review Responses: Identify gaps and strengths.
- Develop Action Plans: Create plans to address gaps.
- Assign Responsibilities: Allocate tasks to team members.

5. Debrief & Feedback

- Conduct a Debrief: Discuss the exercise experience.
- Gather Feedback: Collect feedback for future improvements.

6. Follow Up

- Implement Action Plans: Track and implement plans.
- Schedule Follow-Up Exercises: Plan regular exercises for continuous improvement.

Inject

Reports flood in from your customer service team: multiple users are experiencing failed authentications, receiving incorrect responses to their queries, and getting information about other customers.

1

Questions

- What are your immediate steps to investigate and triage the issue?
- How do you isolate the issue to determine if it's an AI failure, infrastructure problem, or potential attack?

Inject

You receive word that not only are customer queries being mishandled, but also unauthorized transactions have occurred. Some customers report receiving personal information unrelated to their account.

2

Questions

- What is your immediate course of action to prevent further exposure of sensitive information.
- How do you determine if this sensitive information exposure is due to a technical malfunction, user error, or a security breach?

Inject

Your company contacts the AI development team, which reports that the system's model was trained on recent customer interaction data, and initial investigation shows no data processing anomalies. However, the IT team suspects potential unauthorized access to the system, which might have led to the altered behaviour.

3

Questions

- How do you assess whether Apollo's misleading responses are due to data processing errors, insufficient training, or deliberate manipulation?
- How do you ensure users do not inadvertently expose sensitive company data to the AI model?

Inject

More customers report receiving sensitive personal information that doesn't belong to them. Your Legal and compliance teams are concerned about potential data breach laws and GDPR violations.

4

Questions

- What are your immediate actions to comply with legal regulations on data breaches?
- How do you investigate the source of the sensitive information in the AI model's responses?

Inject

A third-party security firm is hired to perform a security audit. Their findings suggest that the AI system may be using a third-party library vulnerable to a prompt injection attack. The firm also indicates that attackers might have been able to manipulate the AI model through these vulnerabilities.

5

Questions

- What steps would you take to address the prompt injection vulnerability?
- How do you validate that the AI model is secure, and what ongoing monitoring do you establish to detect issues?