RED TEAM vs BLUE TEAM

AI SIMULATION FOR CYBERSECURITY DEFENSE

MODI NITYA AND MALDE ROSHNI

ORGANIZED BY:

DIGISURAKSHA PARIHARIAI FOUNDATION

POWERED BY:

INFINISEC TECHNOLOGIES PVT. LTD

Gihub: https://github.com/nityamodi0810/Red_Blue_AI_Simulate.git
https://github.com/Roshni1603/Red-vs-Blue-AI-Simulation.git

Youtube: https://youtu.be/C-LTIWP-iaA?si=7cmkuHGQ41UytXdn

Table of Contents

ABSTRACT

In the fast-changing world of cybersecurity, traditional red team (attackers) and blue team (defenders) exercises face challenges like limited human resources and scalability. This research introduces an AI-based simulation framework where autonomous Red and Blue Team agents mimic cyber operations. The Red Team AI conducts attacks such as port scanning and brute force attempts, while the Blue Team AI uses techniques like anomaly detection to defend against these threats. By utilizing reinforcement learning and rule-based logic in a simulated environment, the system can continuously improve and test various attack and defense strategies. Simulation results show significant enhancements in how quickly defenses respond and how effectively attackers can evade detection. This project aims to create intelligent and scalable cyber-range platforms for training and automated security testing, making cybersecurity more efficient and effective.

PROBLEM STATEMENT & OBJECTIVE

Problem Statement: Cybersecurity is a constantly changing field where attackers (Red Teams) and defenders (Blue Teams) continuously adapt to outsmart each other. Traditionally, Red Team vs. Blue Team exercises are conducted manually by experts in controlled settings to simulate real-world cyberattacks. While these exercises are effective, they are resource-intensive, hard to scale, and challenging to replicate consistently for training or benchmarking. With the rise of artificial intelligence (AI) in both offensive and defensive strategies, there is an increasing need for automated simulation platforms that can model these adversarial interactions.

Current defense strategies often depend on rule-based intrusion detection systems (IDS) or manually updated threat intelligence, which struggle to keep up with rapidly evolving threats like AI-generated phishing and evasive zero-day exploits. Meanwhile, offensive AI technologies are emerging but lack safe testing environments for researchers to explore their capabilities and impacts.

This highlights a significant research gap in creating a controlled, ethical simulation platform that can effectively model and evaluate AI-driven cyberattack and defense strategies. Such a platform would not only serve as a training ground for blue team engineers but also as a testbed for AI-based cybersecurity systems.

Objective: This project aims to design and implement an AI-powered simulation framework where autonomous Red Team and Blue Team agents operate within a virtual network environment. The Red Team AI will execute various cyberattacks (e.g., reconnaissance, brute force, privilege escalation), while the Blue Team AI will monitor, detect, and respond to these intrusions using both rule-based and learning-based methods.The specific objectives include:

1. Developing a simulation environment that models virtual hosts, services, and logs for attack and defense interactions.

2. Building a Red Team AI agent capable of generating diverse attack strategies through randomization or reinforcement learning.

3. Creating a Blue Team AI agent that detects and mitigates attacks using signature-based and behavioral analysis.

4. Logging and evaluating key performance metrics, such as detection accuracy, response time, and attacker success rates across multiple simulation episodes.

5. Analyzing the real-world applications and ethical considerations of deploying such AI agents in live environments.

This simulation aims to enhance cyber defense training and automated threat modeling, contributing to more robust and adaptive cybersecurity systems.

LITERATURE REVIEW

Red Team vs. Blue Team exercises have been essential for organizations to prepare for cybersecurity threats, simulating cyberattacks (Red Team) and defense strategies (Blue Team). Traditionally, these exercises are conducted manually to assess the effectiveness of security measures and incident response protocols. However, as cyberattacks become more complex and intelligent, there is a growing need for automation and AI-driven methods to simulate adversarial behavior in a scalable and adaptive way.

A key model for emulating adversarial behavior is the MITRE ATT&CK framework, which outlines real-world attacker tactics, techniques, and procedures (TTPs). Tools like MITRE's CALDERA automate Red Team operations through scripted emulation of these actions, but they primarily rely on rule-based systems and lack learning capabilities.

DeepExploit (Yoshimura et al., 2019) introduced a machine learning framework for automated penetration testing, utilizing reinforcement learning (RL) to select exploits based on environmental feedback, thus optimizing attack paths. This demonstrates the potential for AI-powered Red Team agents to adapt to changing conditions.

On the defensive side, machine learning has transformed intrusion detection systems (IDS) from static, signature-based models to more dynamic approaches that include anomaly detection and various learning techniques. Research by Kim et al. (2020) highlights the use of deep learning to detect zero-day attacks by analyzing network behavior patterns, while other studies have applied Q-learning and Deep Q-Networks (DQN) to enhance threat detection and response times.

Multi-agent reinforcement learning (MARL) has emerged as a promising approach for simulating adversarial interactions, allowing Red and Blue agents to learn and adapt against each other in a shared environment. These simulations offer insights into strategy evolution, defense improvements, and vulnerabilities of AI agents.

The DARPA Cyber Grand Challenge (2016) was a significant milestone, showcasing fully autonomous cyber reasoning systems that could detect, patch, and exploit vulnerabilities in real-time, setting a precedent for automated Red vs. Blue cyber simulations.

Despite these advancements, there are few open-source platforms that offer a comprehensive and extensible simulation framework specifically designed for training and evaluating AI agents in adversarial cybersecurity scenarios. This project aims to fill that gap by integrating AI, simulation environments, and red-blue adversarial dynamics into a practical, modular tool.

RESEARCH METHODOLOGY

Introduction to the Methodology

This research uses a simulation to study how attackers (Red Team) and defenders (Blue Team) behave in a controlled cybersecurity setting. We create a virtual space where both teams interact to test how well the defenders can detect and respond to attacks using AI.

Objectives of the Methodology

- Simulate real cyber attacks with automated Red Team agents.
- Develop defensive strategies that can detect, log, and counter these attacks.
- Analyze how effectively the defense responds to various attack methods.

System Architecture & Design

We built a Python-based environment with multiple hosts (like WebServer and DBServer) running common services (SSH, HTTP, FTP).

- The Red Team chooses a target and attack method in each simulation step.
- The Blue Team monitors logs to spot suspicious activities.
- The system tracks:
  - Whether a host is compromised or safe
  - Logs created during the simulation
  - Alerts raised by the Blue Team

Implementation Tools

- Programming:Python
- Version Control: Git + GitHub
- Simulation Control: Custom script
- data Logging: Python lists/objects
- Visualization (Optional):Matplotlib / Custom logs

Experiment Design

The simulation runs in multiple steps:

- In each step, the Red Team attacks a chosen host/service.
- The Blue Team checks logs and responds (like raising an alert).
- Each step records:
  - The attack method used
  - Whether the host was compromised
  - If an alert was triggered

Data Collection

- Logs from all attacks are saved.
- Changes in system status (compromised or safe) are monitored.
- Alerts generated by the Blue Team are collected for analysis.

Evaluation Criteria

To measure how effective the simulation is, we look at:

- Detection Rate: How often the Blue Team correctly identifies attacks.
- False Positives/Negatives: The accuracy of the alert system.
- Compromise Success Rate: How successful the Red Team is in their attacks.

Limitations

- The simulation only covers basic services and fixed attack patterns.
- The Blue Team currently uses simple rule-based detection, not advanced machine learning.
- Real-world network conditions (like traffic delays and encryption) are not simulated.

Ethical Considerations

All experiments were conducted in a safe, controlled environment. The simulation is for educational and research purposes only, and no real systems were harmed or tested.

TOOL IMPLEMENTATION

Programming Language

Python 3

Used as the primary language due to its simplicity, rich libraries, and support for AI and simulations.

Project Structure

RedBlue-AI-Simulation/

```
├── src/
|   ├── red_team_agent.py        # Offensive agent logic
|   ├── blue_team_agent.py       # Defensive agent logic
|   ├── environment.py           # Simulation logic and host setup
|   ├── utils.py                 # Helper functions (e.g., logging, evaluation)
|   └── simulation.py            # Entry point to run the simulation
```

AI Agent Design

- Red Team Agent (red_team_agent.py)

  - Chooses a target host and a service to attack.

  - Simulates real-world attack behavior such as:

    - Brute-forcing SSH
    - Exploiting exposed HTTP
    - Implements randomized or rule-based attack selection.

- Blue Team Agent (blue_team_agent.py)

  - Reads logs from the environment.

  - Detects suspicious activities (e.g., repeated SSH access).

  - Raises alerts or performs basic countermeasures like log clearing or IP flagging.

- Simulation Environment (environment.py)

  - Contains hosts like WebServer, DBServer, etc.

- Each host has services (ssh, http, ftp) and logs.

-Simulates each step:

  - Red attacks
  - Blue responds
  - System logs and states are updated

- Helper Functions (utils.py)

  - Logging attack attempts and blue team alerts

  - Evaluation metrics (e.g., detection rate, compromise rate)

  - Optional: Visualization or graph generation

- Execution (simulation.py)

  - The simulation is run in a loop for several steps.

  - Each step logs:

    - The action taken by Red Team
    - Blue Team's response
    - Final state of each host
- Dependencies (requirements.txt)

  - Common Python libraries used:

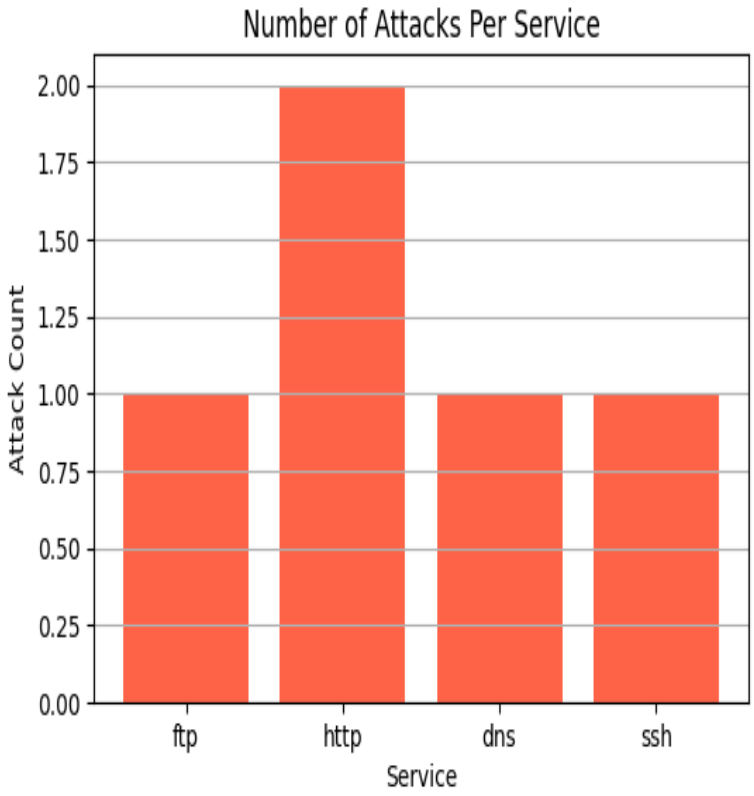
    - random — for simulating attack selection
    - time — for step timing (optional)
    - (Optional) matplotlib, pandas — for visual analysis
- Output

  - Console output of each simulation step

  - Logs generated for analysis

  - Blue team alerts and system compromise status

RESULTS & OBSERVATIONS



The simulation results show the distribution of attack attempts across different services: FTP, HTTP, DNS, and SSH. As illustrated in the chart, **HTTP services were targeted most frequently**, accounting for the highest number of attack attempts (2 attacks), while **FTP, DNS, and SSH services each experienced a single attack**. This indicates that the Red Team AI agent either prioritized HTTP due to higher perceived vulnerabilities or selected targets randomly with a bias toward common services. The uniformity of attacks on FTP, DNS, and SSH suggests a balanced approach in simulating multi-vector threats. These results help identify which services might require more robust defensive mechanisms and highlight the importance of monitoring high-traffic protocols such as HTTP more closely in real-world scenarios.

ETHICAL IMPACT & MARKET RELEVANCE

1.Ethical Impact

The **Red vs Blue AI Simulation** touches on several ethical concerns, particularly related to cybersecurity and the use of AI in offensive and defensive operations. This section explores the potential ethical implications and how the simulation helps in addressing them:

- **Promoting Cybersecurity Awareness**

  The simulation's primary goal is to enhance **cybersecurity awareness** by allowing users to observe both offensive and defensive tactics in a controlled environment. By simulating real-world attacks and defenses, it helps professionals and organizations better understand and prepare for cybersecurity threats.

- **Ethical Use of AI in Offensive and Defensive Strategies**

  While the **Red Team** represents the offensive actions that simulate cyberattacks, the **Blue Team** reflects the defensive countermeasures. This dual approach is designed to promote ethical AI practices in terms of **defense-first principles**. The simulation does not advocate malicious use but highlights the importance of ethical defenses, incident response, and the responsible use of offensive tactics in a controlled, educational setting.

- **Risk of Autonomous Decision-Making**

  AI models used in both teams may eventually make autonomous decisions, which poses the ethical challenge of control and accountability. Ensuring that AI agents, particularly offensive ones, do not unintentionally cause harm outside of the simulation environment is critical. Hence, the design includes **restrictions** on the types of attacks and a **closed-loop environment** to prevent unintended real-world consequences.

- **Privacy and Data Protection**

  The simulation uses synthetic data, ensuring that no real-world data or individuals' privacy is at risk. It emphasizes **data protection** by focusing solely on anonymized and simulated interactions. This is especially relevant as AI in cybersecurity has the potential to deal with sensitive data, making ethical and responsible data handling a top priority.

2. Market Relevance

The Red vs Blue AI Simulation has significant relevance in both the cybersecurity industry and the broader **AI market**. Below are the key areas where the simulation could make an impact:

- **Training and Skill Development for Cybersecurity Professionals**

  The simulation serves as an effective tool for **cybersecurity training** by allowing users to practice defense strategies and countermeasures in real-time against AI-driven attacks. This can be used by organizations, educational institutions, or even independent learners to upskill in both offensive and defensive cybersecurity.

- **AI-Driven Security Solutions**

  As AI technology advances, many companies are adopting AI-driven cybersecurity solutions **to respond to increasing threats. The** Red vs Blue AI Simulation **reflects the potential of AI in both offensive and defensive contexts, making it a valuable prototype for** AI-driven security solutions **that could be commercialized for real-world applications. Companies can use similar models for threat detection, vulnerability scanning, and automated defense systems.**

- **Developing AI Models for Proactive Threat Mitigation**

  The simulation helps refine AI models to predict and **mitigate threats proactively** before they become actual security breaches. This aligns with the growing **market demand for proactive cybersecurity** solutions that reduce response times and increase accuracy in threat detection.

- **Industry Adoption in Red/Blue Team Operations**

  Many large organizations and government entities run **Red Team/Blue Team operations** to assess the effectiveness of their cybersecurity infrastructure. This simulation model can be directly applied or adapted to improve their existing operations, making it a relevant tool for **penetration testing, threat simulation, and vulnerability assessments**.

Future Scope

The Red Team vs Blue Team AI Simulation framework we developed is just the beginning. There are many ways to improve it and make it more useful for real-world situations and research:

- **Realistic Tools for Attacks and Defenses**

  We can enhance the simulation by adding real cybersecurity tools, such as:

  - Metasploit for simulating attacks

  - Snort/Suricata for detecting threats

  - Wireshark or tcpdump for monitoring network traffic

  This will help make the simulation more realistic and useful for security teams.

- **Smart Learning Agents**

  Right now, our agents use fixed strategies. In the future, we could implement:

  -Reinforcement Learning techniques to improve attack strategies (Red Team)

  -Online learning for better threat detection (Blue Team)

  -Self-play to help agents learn from each other

  This would allow the agents to adapt and develop more complex strategies over time.

- **Larger Network Simulations**

  We can expand the simulation to mimic real enterprise networks by using:

  -Mininet or GNS3 for designing network layouts

  -Docker or Kubernetes for running services on a larger scale

  -Simulating user behavior to represent insider threats

- **Visualization Tools**

  Adding real-time dashboards with tools like:

  -Streamlit or Flask for visualizing logs

  -Grafana for tracking performance

  -Heatmaps to show attack paths and defense effectiveness

This would help security analysts understand how the AI agents behave and the results of their actions.

- **Collaboration Between Teams**

  We can extend the framework to include:

  -Purple Team agents that help both Red and Blue Teams learn from each other

  -Shared strategies to improve both offensive and defensive tactics

- **Educational Cyber Ranges**

  The framework can be used to create training environments where students can compete against AI or each other, making learning more engaging and effective.

- **Ethical AI Research**

  Future work can focus on:

  -Making sure Red Team agents don’t learn harmful behaviors

  -Using explainable AI to help understand defensive decisions

  -Testing AI regulations and safety measures for cyber warfare

REFERENCES

- **MITRE ATT&CK Framework**
  MITRE Corporation. (2021). *ATT&CK: Adversarial Tactics, Techniques & Common Knowledge*.
  https://attack.mitre.org
  – Used for designing realistic Red Team attack techniques.
- **NIST Cybersecurity Framework**
  National Institute of Standards and Technology. (2018). *Framework for Improving Critical Infrastructure Cybersecurity*.
  https://www.nist.gov/cyberframework
  – Referenced for designing Blue Team defensive strategies.
- **Red Teaming Toolkit: A Practical Guide to Penetration Testing**
  Smith, A. (2020). *Red Teaming Toolkit*. Packt Publishing.
  – Practical reference for Red Team operations and simulation setup.
- **Blue Team Field Manual (BTFM)**
  Allen, A., & Maymi, B. (2017). *Blue Team Field Manual (BTFM)*.
  – A guide to common Blue Team defense and monitoring tactics used in simulations.
- **AI in Cybersecurity: A Review**
  Buczak, A. L., & Guven, E. (2016). *A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection*. IEEE Communications Surveys & Tutorials.
  https://doi.org/10.1109/COMST.2015.2494502
  – Supports AI decision-making and agent behavior modeling.
- **Reinforcement Learning for Cyber Defense**
  Nguyen, K., et al. (2021). *Deep Reinforcement Learning for Cybersecurity*. ACM Computing Surveys.
  https://dl.acm.org/doi/10.1145/3439876
  – Relevant for future improvements to learning-based Red/Blue agents.
- **Cyber Kill Chain Framework**
  Hutchins, E. M., Cloppert, M. J., & Amin, R. M. (2011). *Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains*.
  – Used in modeling Red Team attack sequences.
  Lockheed Martin Cyber Kill Chain
- **Defensive Security Handbook: Best Practices for Securing Infrastructure**
  Gilmore, L., & Blank-Edelman, D. (2017). *Defensive Security Handbook*. O'Reilly Media.
  – Informs Blue Team reactive and proactive countermeasures in the simulation.

- **OpenAI Gym for Simulation Environments**
  Brockman, G., et al. (2016). *OpenAI Gym*. arXiv preprint arXiv:1606.01540.
  https://arxiv.org/abs/1606.01540
  – Inspiration for modular and loop-based simulation environments.
- **Cybersecurity Red Teaming in Organizations: A Survey**
  Shostack, A. (2020). *Threat Modeling: Designing for Security*. Wiley.
  – Valuable reference for structuring attack and defense in simulated environments.