Abstract:
Artificial Intelligence (AI) has helped cybersecurity through the ability to analyze and evaluate massive amounts of data and discover patterns at far faster rates than humans can. However, AI cannot replace the involvement of humans within the cybersecurity AI has specific limitations: it cannot understand the contextual meaning of information; identify new attacks; make ethical decisions, nor create ideas. The reason for this limitation is because the security architecture was created before the existence of AI technology. Consequently, there has been a systematic issue of the lack of a standardized methodology by which AI can be integrated into the system of cybersecurity. Traditional frameworks do not contain comprehensive descriptions of the traditional methods of governance of the probabilistic models created by AI. Therefore, to place total reliance upon autonomous systems to perform critical security functions is both inappropriate at this time, as well as unacceptably high-risk. The optimal solution for cyber defense is a hybrid model in which humans and AI complement each other's weaknesses.

Introduction:

The present research addresses the rapid technological change within the cybersecurity context, where the integration of AI and ML emerges as the most prominent development over the last ten years.

Increasing investment in the AI cybersecurity market, due to expectations related to automation and decreasing human error, is projected to exceed $46 billion by 2027.

Persistent threats are still there: despite considerable investments and advanced tooling, cyber-attacks with successful data breaches still happen.

Limitations of AI are evident from the continuing security breaches in organizations with AI-enhanced defenses, showing that AI alone is not enough and has fundamental constraints when it comes to handling complex threat landscapes.

Hypothesis: AI will not supplant human cybersecurity experts but rather serve as one more strong tool.

Critical limitations identified include shortfalls in:

- Contextual understanding
- Handling Novel, Unseen Attacks: Zero-day Threats
- Ethical and business judgment
- Creative, adversarial problem-solving (hacker-like)
- Accuracy in handling false positives and false negatives

Problem Statement:

The key research question is: Can reliance on AI for cybersecurity operations be complete-in the particular context of SOC, penetration testing, and vulnerability assessment?

The working hypothesis is that, while AI offers substantial value during cybersecurity operations, it cannot fully replace human judgment because of certain core limitations it suffers from in:

1) Understanding context and nuance;
2) adapting to novel, zero-day threats;
3) Making ethical and business-critical decisions
4) use creative problem solving and think laterally
5) determining adversarial intent and motivation
6) Managing false positives and false negatives effectively.

Research Objectives

This study aims to:

1. Present the current landscape of AI in deployment within Security Operations Centers - SOC, Penetration Testing, and Vulnerability Assessment.
2. Identify specific limitations and failure modes of AI across these domains.
3. Document real case studies about failures of AI-based security solutions.
4. Propose frameworks to optimize human–AI collaboration in cybersecurity.
5. Recommend best practices for organizations employing AI in security operations.

Scope and Methodology

The study focuses on three key cybersecurity domains:

- **Security Operations Centers (SOC):** Real-time monitoring, incident detection, and response.
- **Penetration Testing:** Offensive security testing and ethical hacking.
- **Vulnerability Assessment:** This is the process of identifying, classifying, and prioritizing the security weaknesses of an organization.

The methodology used includes

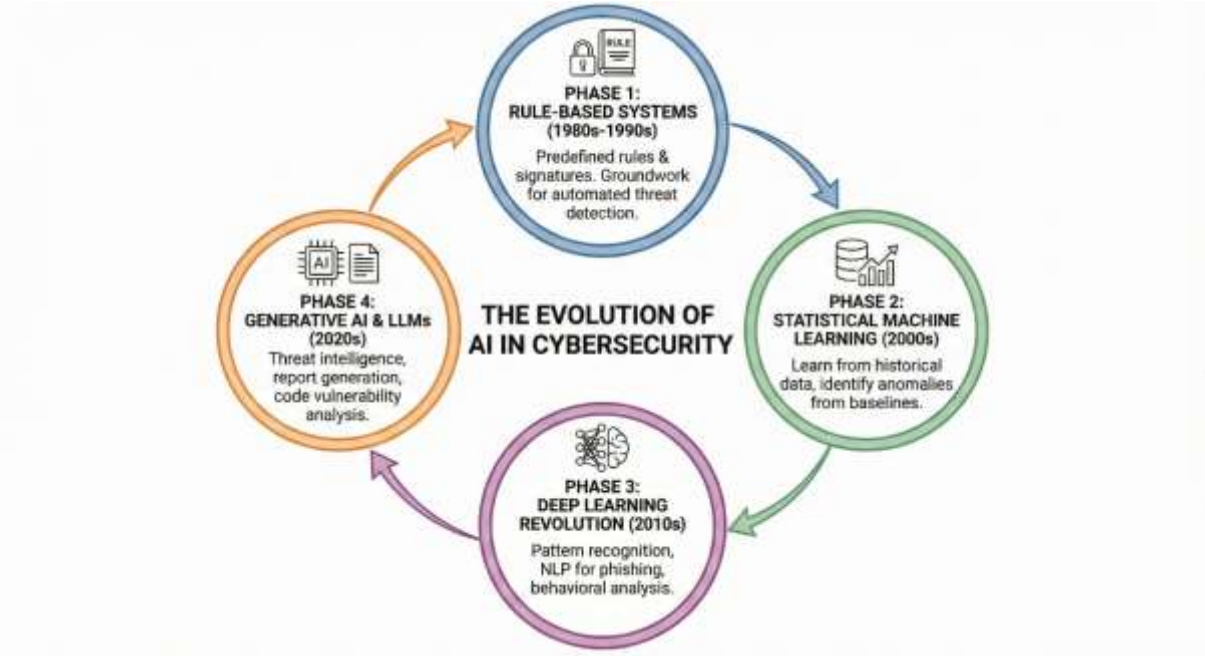
- A systematic literature review of scholarly publications and industry reports.
- Analysis of documented AI failures in cybersecurity contexts.
- Expert interviews and practitioner insights.
- Comparative analysis of AI performance versus human performance metrics.

Literature Review

Evolution of AI in Cybersecurity

The application of AI to cybersecurity has evolved through several distinct phases:

- **Phase 1:** Rule-Based Systems (1980s-1990s) Early intrusion detection systems relied on predefined rules and signatures. These systems, while not "AI" in the modern sense, laid the groundwork for automated threat detection.
- **Phase 2:** Statistical Machine Learning (2000s) The introduction of statistical methods enabled systems to learn from historical data, identifying anomalies based on deviation from established baselines.
- **Phase 3:** Deep Learning Revolution (2010s) Neural networks and deep learning enabled more sophisticated pattern recognition, natural language processing for phishing detection, and behavioral analysis.
- **Phase 4:** Generative AI and Large Language Models (2020s) The emergence of LLMs introduced new capabilities in threat intelligence analysis, automated report generation, and code vulnerability analysis.



Current AI Capabilities in Cybersecurity

Contemporary AI systems in cybersecurity demonstrate proficiency in:

Capability	Effectiveness	Limitations
Pattern Recognition	High	Limited to known patterns
Anomaly Detection	Medium-High	High false positive rates
Malware Classification	High	Struggles with polymorphic malware
Phishing Detection	Medium-High	Evaded by sophisticated attacks
Log Analysis	High	Context-dependent accuracy
Automated Response	Medium	Risk of automated errors
Threat Intelligence	Medium	Requires human validation

The Human Factor in Cybersecurity

The empirical evidence does suggest that human expertise adds irreplaceable value across a number of domains:

1) Contextual Understanding: Humans understand organizational context, business priorities, and acceptable levels of risk.
2) Creative Thinking: Security professionals are able to adopt attacker mind-sets to predict new attack vectors.
3) Ethical Decision-Making: Humans solve complex ethical dilemmas arising in security responses.
4) Relationship Building: Human analysts work well with stakeholders and communicate difficult findings.
5) Adaptability: Humans quickly respond and adapt to unprecedented situations.

Overview of AI in Security Operation Center (SOC) Operations

AI is increasingly integrated into modern Security Operations Centers to further improve the following:

- Security Information and Event Management (SIEM) functions
- User and Entity Behaviour Analytics - UEBA
- Automation of Alert Triage and Prioritization
- Threat Intelligence Correlation
- Automation of incident response

Critical Limitations

The False Positive Problem

Problem: AI-powered detection systems generate a high level of false positives, thereby causing alert fatigue among human analysts.

Statistics

- The average SOC deals with more than 11,000 alerts per day.
- 52% of alerts are false positives. SANS Institute, 2023
- Analysts spend around 25 percent of their time searching for false alarms

Why AI Fails:

- Machine learning models optimize detection performance, often at the cost of precision.
- Legitimate yet atypical behavior triggers anomaly detection
- Activities that depend on context may be incorrectly classified as malicious without proper understanding of the context

Adversarial Machine Learning Vulnerabilities

Issue: Adversaries may craft attacks to deliberately evade or manipulate AI-based detection systems.

Attack Types:

1) Evasion Attacks: Modifying malicious payloads to evade detection.
2) Poisoning Attacks: Compromising training data to introduce backdoors.
3) Model Extraction: Exfiltrating ML model parameters to understand the underlying detection logic.
4) Model Inversion: Inferring sensitive training data from model outputs.

Impact: This puts organizations relying exclusively on AI-based detection at increased risk from adversarial attacks crafted specifically to exploit weaknesses in machine learning systems.

Automated Response Risks

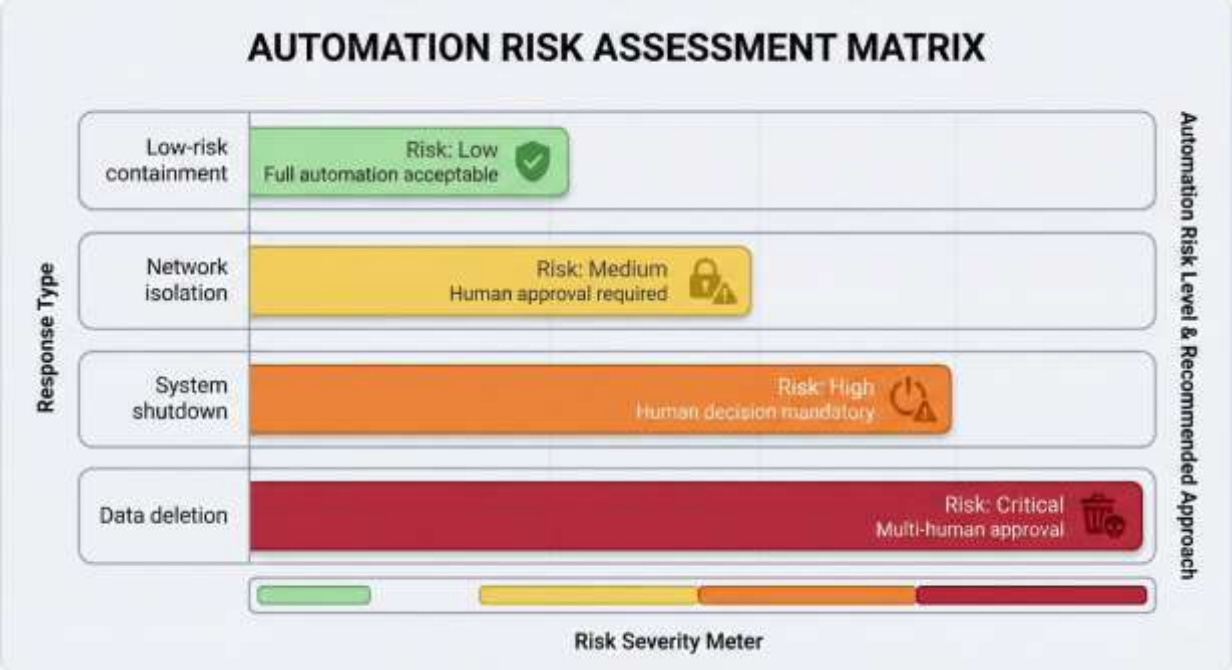
Problem: AI-driven automated responses, if triggered inappropriately, can cause operational disruption.

Documented incidents:

- Automated systems quarantining legitimate business applications
- Network segmentation due to false positives that trigger outages.
- Automated account lockouts affecting executives during critical negotiations.

Risk Matrix:

Response Type	Automation Risk Level	Recommended Approach
Low-risk containment	Low	Full automation acceptable
Network isolation	Medium	Human approval required
System shutdown	High	Human decision mandatory
Data deletion	Critical	Multi-human approval



Limitations of AI Replacement in SOC

1) Threat Hunting: Proactive, hypothesis-driven searches require human ingenuity.
2) Incident Command: Human leadership and decision-making authority are needed for the management of critical incidents.
3) Stakeholder Communication: Communicating incident information to executives relies upon human communication skills.
4) Root Cause Analysis: Elucidation of root causes requires contextual reasoning.
5) Policy Decisions: Judging acceptable risk depends upon human judgment.

Overview of AI in Penetration Testing

AI tools used in penetration testing include:

- Automated vulnerability scanners
- Intelligent exploitation frameworks
- Password cracking optimization
- Automation of social engineering attacks
- Tools to create technical reports

Critical Limitations

Lack of Creative Exploitation

Problem: AI cannot mimic the creative, lateral thinking that comes courtesy of seasoned penetration testers.

Human Advantages:

- Ability to chain multiple low-risk vulnerabilities into impactful exploits
- Knowledge of business logic vulnerabilities Innovative approaches to social engineering
- Consideration of out-of-scope attack paths

Physical Security Testing

Problem: Artificial intelligence cannot do physical penetration testing, which is a domain where many critical attack vectors are seen regularly.

Improving Physical Testing: Elements of Physical Security Assessment

1) Tailgating to access secured facilities
2) Physical social engineering
3) Badge Cloning and RFID-based Exploits
4) USB Drop Attack Scenarios
5) Dumpster diving for sensitive information
6) Physical lock bypass techniques
7) Surveillance and reconnaissance activities

Impact: An organization that depends solely on automated system testing may miss critical gaps in physical security.

Social Engineering

This domain is quite challenging because effective social engineering requires subtle knowledge about psychology, culture, and context. The limitations of artificial intelligence in this domain include:

- Pretext development demands cultural literacy
- Human judgment for real-time adaptation in telecommunication interactions
- Rapport-building is essentially a human capability and relies on emotional intelligence to pick up on emotional cues.

AI vs. Human Capability Comparison

Technique	AI Capability	Human Capability
Mass phishing emails	High	Medium
Spear phishing	Medium	Very High
Vishing (voice phishing)	Low	Very High
In-person pretexting	None	Very High
Relationship building	None	Very High

Ethical Decision-Making

Argument: The testing for penetration requires constant ethical judgment, an ability artificial intelligence lacks.

Ethical Decisions Requiring Human Judgement

1) Determining when testing should stop to prevent substantive harm
2) Managing the Discovery of Illegal Activities
3) Handling Discovered Vulnerabilities Affecting Third Parties
4) Protecting Discovered Sensitive Information
5) Evaluating Potential Impacts on real users during testing

What AI Cannot Replace in Penetration Testing

1) Attack Creativity: Identification of new attack vectors
2) Human Targets: Social engineering involving human interaction
3) Physical Testing: All components of physical security assessment
4) Ethical Judgment: Real-time ethical decision making
5) Strategic Evaluation: Understanding of the organizational context
6) Client Communication: Building trust and presenting results

Overview of AI in Vulnerability Assessment

Artificial intelligence is used in vulnerability assessment for:

1) Automated scanning and discovery
2) Vulnerability Prioritization and Scoring
3) Patch recommendation engines
4) Risk assessment models
5) Continuous monitoring

Critical Limitations

Issue: AI does vulnerability prioritization purely from a technical standpoint, without consideration for business context.

The problem:

- Only the CVSS scores do not reflect actual organizational risk; critical business assets may not be identifiable by AI
- Compensating controls might be invisible to automated tools
- Business impact varies significantly across organizations

Vulnerability	**CVSS Score**	**AI Priority**	**Actual Priority**	**Reason**
RCE in legacy system	9.8	Critical	Low	System is air-gapped
XSS in customer portal	6.1	Medium	Critical	Direct customer impact
SQLi in internal tool	8.1	High	Medium	Only accessible to trusted users

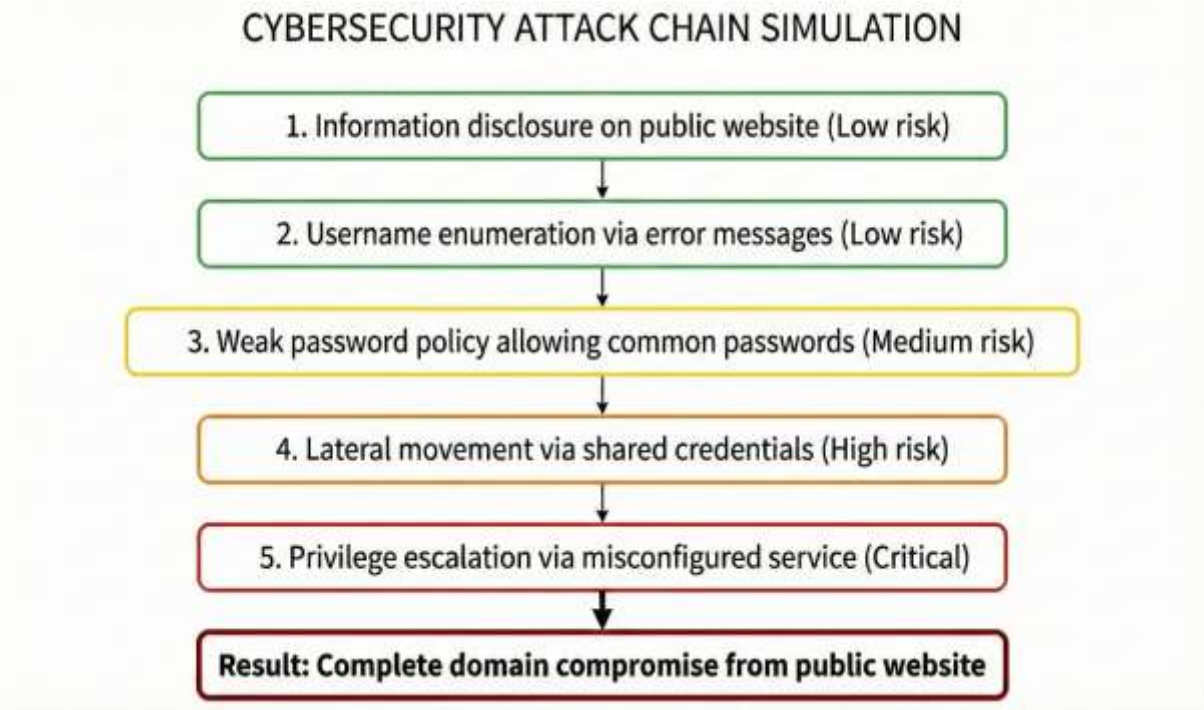
Defining the Ecological Outcomes of Restoration

Problem: Current AI systems are not capable of identifying or inferring business implications related to remediation activities. Many factors that may be elusive to AI's detection include:

- Patches integration into the current systems
- Disruption to business operation due to loss of system functionality
- Availability of personnel possessing relevant skill sets.
- Budgetary constraints and resource limitations
- Regulatory and compliance responsibilities
- Change management procedures

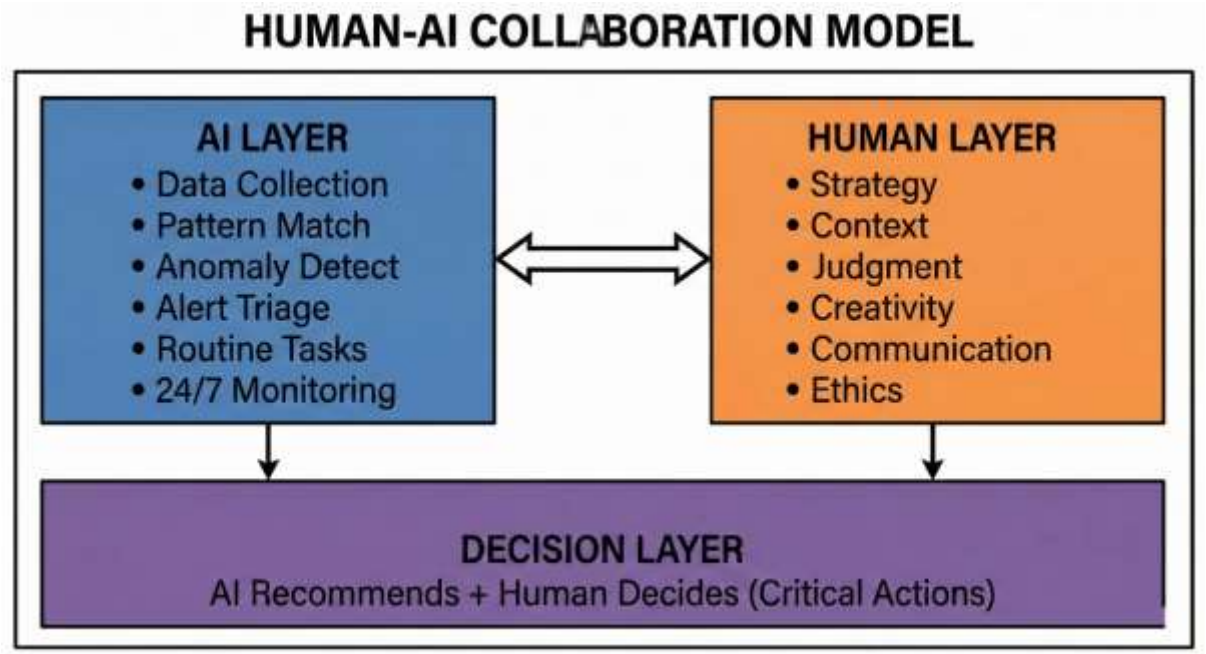
Unrecognized Vulnerability Chaining Problem:

Artificial intelligence is incapable of identifying a sequential combination of low-risk vulnerabilities that masse would present a high-risk attack vector. Rationale and Implications: Individually, the vulnerabilities would be rated as low risk but could, when used in a specific sequence, result in a successful attack. This is a function of AI's tendency to consider vulnerabilities in isolation, while human analysts are better at understanding how one vulnerability might lead to another being exploited.



Limitations of Artificial Intelligence in Vulnerability Assessment

- **Risk in Context Business:** Impact Assessment
- **Remediation Planning:** Balancing Security Imperatives with Operational Continuity
- **Attack Path Analysis:** chains of interdependent vulnerabilities can be identified
- **Compliance Interpretation:** Understanding the Complexities of Regulations and Requirements
- **Stakeholder Communications:** Communicating Risk to Non-Technical Stakeholders
  Prioritization: Align Security Initiatives with Business Objectives



Tier	AI Role	Human Role	Example
Tier 0	Full automation	Audit only	Known-bad IP blocking
Tier 1	Triage & recommend	Review & approve	Phishing email quarantine
Tier 2	Investigate & present	Analyze & decide	Potential insider threat
Tier 3	Support only	Lead investigation	APT activity

Penetration Testing Collaboration Recommendations

Recommended Distribution:

Testing Type	AI Allocation	Human Allocation
Infrastructure scanning	80%	20%
Web application testing	50%	50%
Business logic testing	10%	90%
Social engineering	0%	100%
Physical security	0%	100%
Report writing	30%	70%

Vulnerability Assessment Collaboration Recommendations

Improved Prioritization Model

Recommendation: Make use of AI in determining technical severity scores, but have human judgment involved in the final prioritization.

Scoring Model:

Final Priority = AI Technical Score × 0.4 + Human Business Impact × 0.3 + Human Contextual Factors × 0.3