The 7th IEEE International Conference on Pattern Analysis and Intelligent Systems

The Power of AI in Offensive Cybersecurity: A New Frontier

Dr Mohamed Chahine GHANEM, SFHEA CISSP FCIISec
Associate Professor
Director- Cyber Security Research Centre
London Metropolitan University & University of Liverpool

Laghouat, 23 April 2025

About me

- Engineering Degree (EMP), MSc in Digital Forensics and PhD in Cyber Security Engineering from City, University of London
- 15+ years in the Cyber Security Industry
- Certified Expert (CISSP, CPCI, Multiple SANS GIAC certificates)
- Currently Associate Professor, Director of the Cyber Security Research Centre at London Metropolitan University and Associate Professor at the University of Liverpool
- Chief Consultant in Cyber Security by Design (CSbD) in Banking Sector (prev. Associate Director in Cyber Risk Auditing at Kroll LLC)

Agenda

- ❑ Offensive Cyber Security – Penetration Testing Background
- ❑ AI in Offensive Cyber Security
- ❑ Machine Learning & Deep-Learning Techniques
- ❑ Generative Adversarial Networks & Reinforcement Learning
- ❑ Large Language Models
- ❑ The Cyber Kill Chain & Moving Target Defence
- ❑ Case study
- ❑ Integration of AI with Industry Tools
- ❑ Challenges & Future Directions

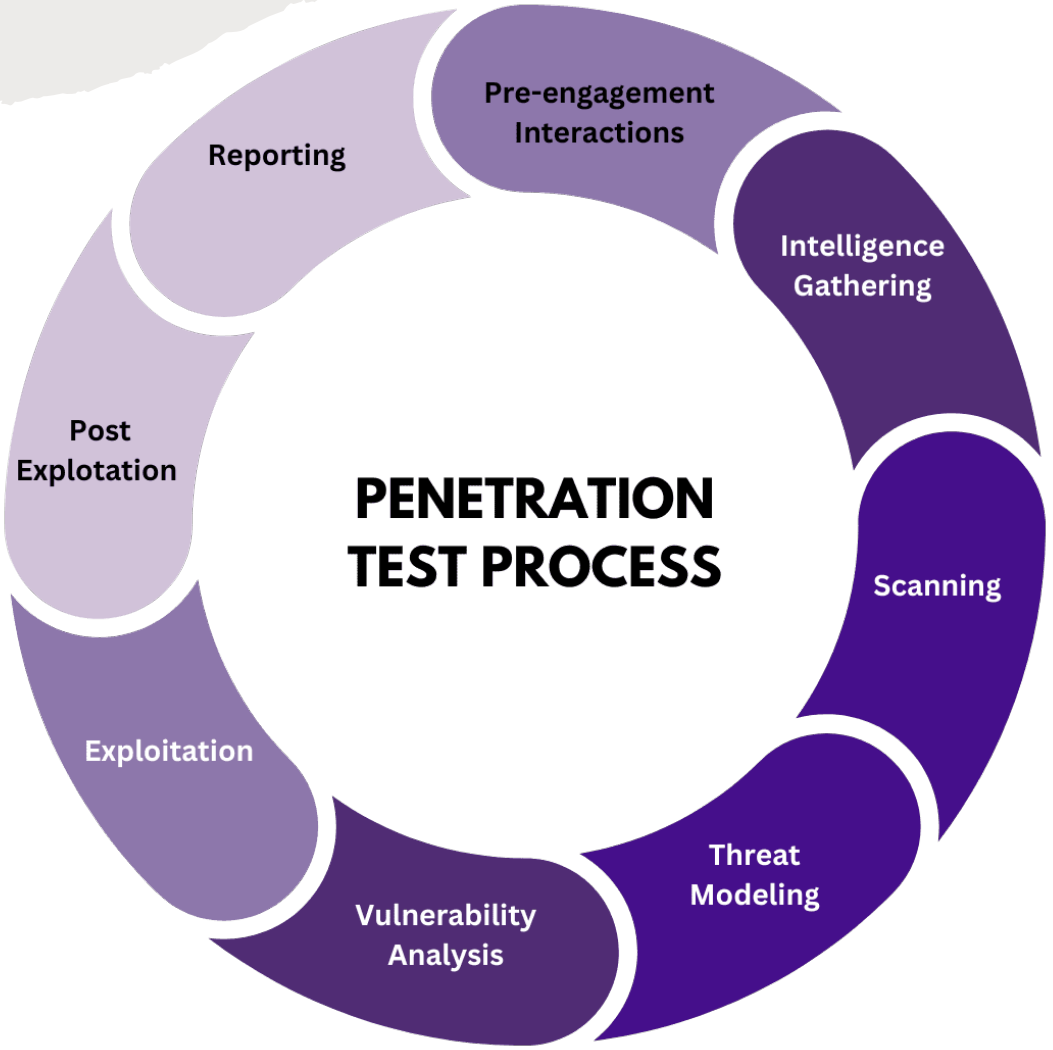
Offensive Security- Penetration Testing

Purpose & Scope: A critical cybersecurity practice that identifies and mitigates vulnerabilities in systems, networks, devices, and applications.

Risk Insights: It evaluates security controls, uncovers risks, and prioritizes improvements to strengthen defences against attacks.

Data Protection: Proactively addresses weaknesses to safeguard sensitive data and prevent breaches.

Ethical Execution: Conducted by skilled "ethical hackers" who simulate real-world attacks to test security resilience.



Why Penetration Testing ? and How ?

1. Detect Vulnerabilities **Before** Criminals Do
2. Test the **Abilities** of IT Asset Defence
3. Assess the **Potential Damage** of a Successful Attack
4. Prove Security **Effectiveness** and **Compliance**
5. Reduce Remediation **Costs** and **Downtime**

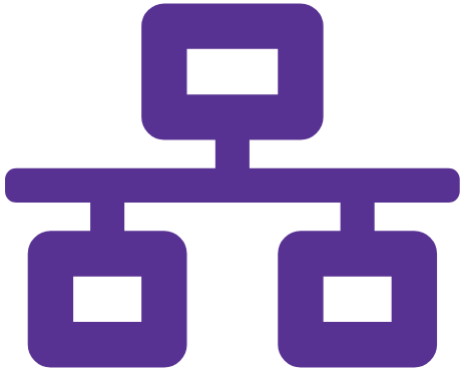
Black Box Testing

Gray Box Testing

White Box Testing

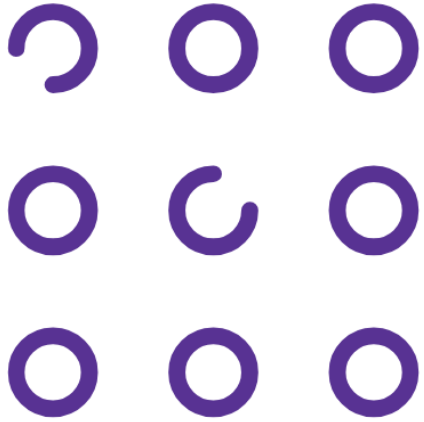
Penetration Testing Scope

Networks and ICT Infrastructures



- External Network
- Internal Network
- Wireless
- Mainframe
- Industrial/SCADA/IIoT

Web Applications



- Web Application
- Mobile Application
- Web Services/API
- App Security Code Review

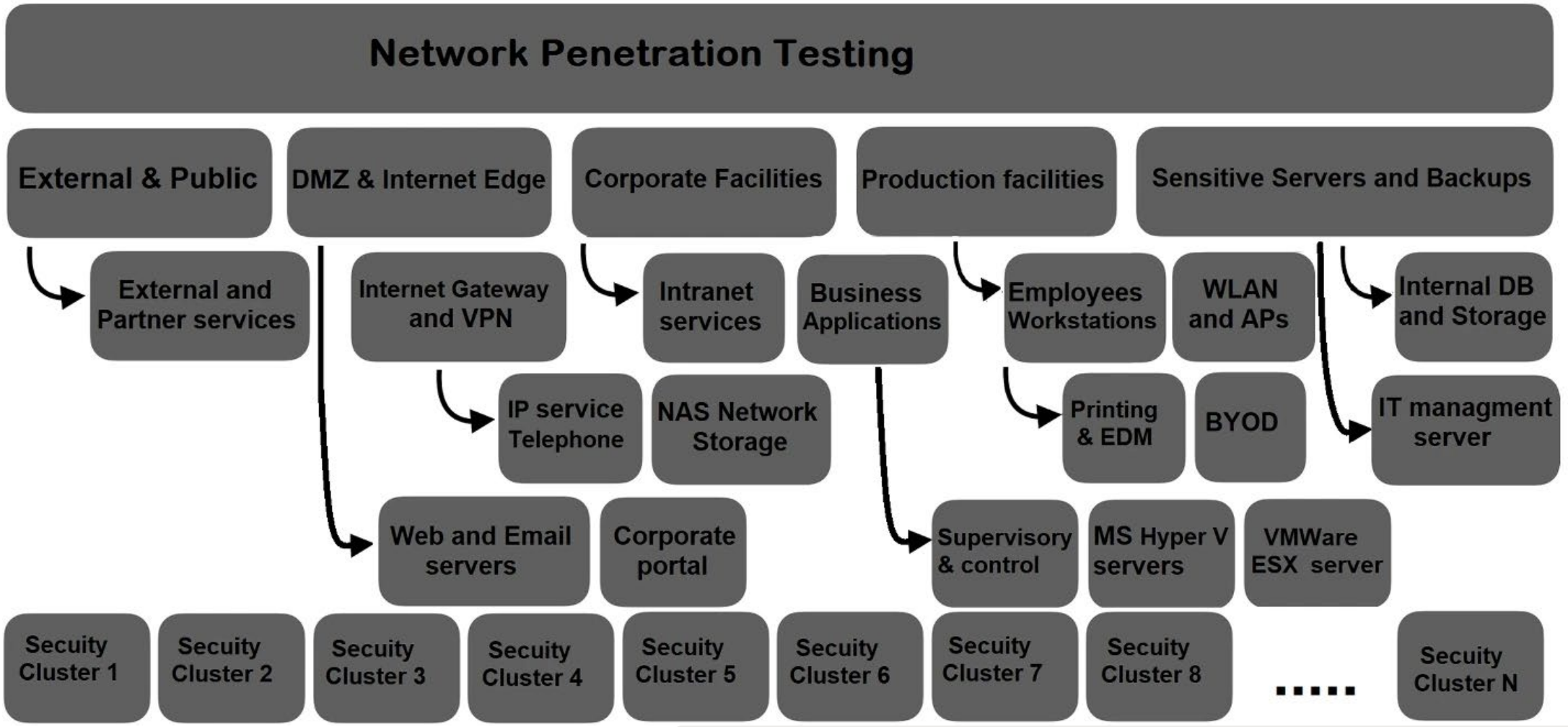
Cloud Services

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)
- Microsoft 365 Security Audit

PT and Security and Compliance

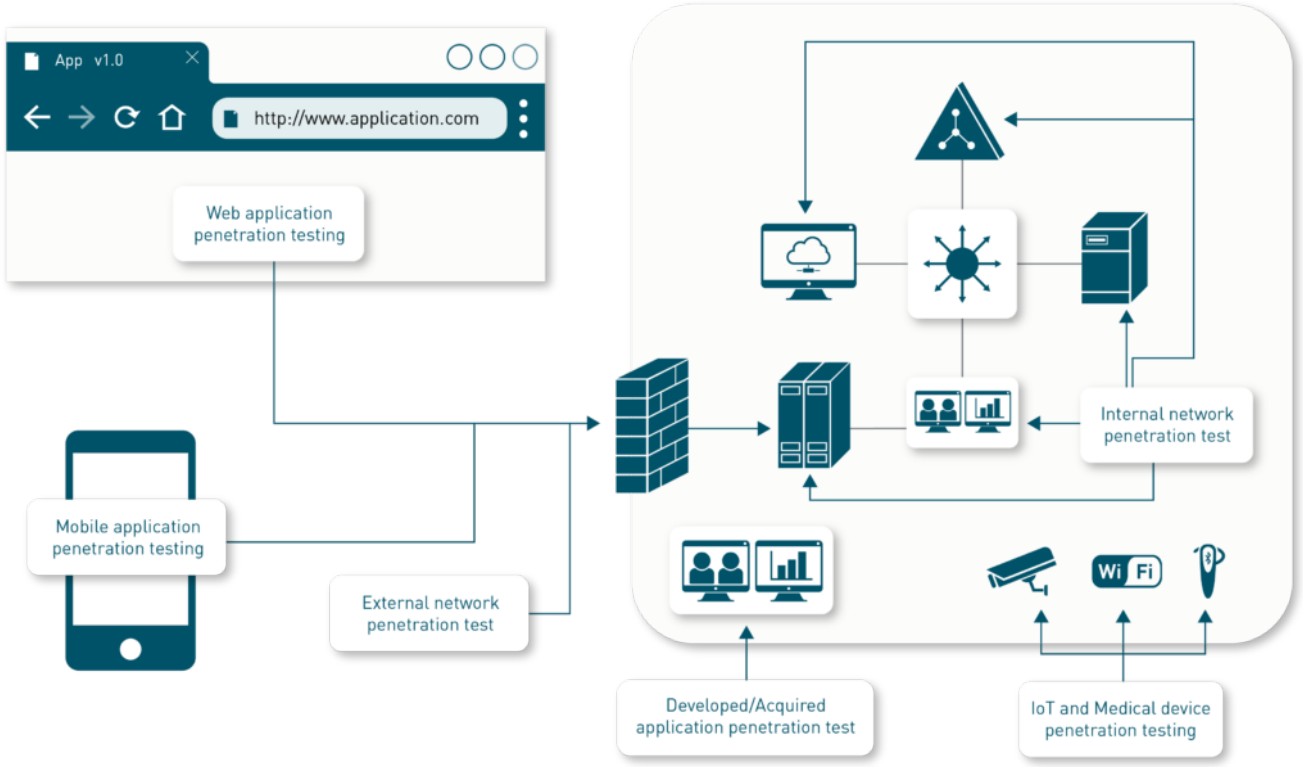
- **ISO27001 ISMS**
- **Payment Card Industry Data Security Standard (PCI DSS)**
- **GDPR Compliance**
- NHS DSP Toolkit
- **SWIFT CSP**
- FDA medical device penetration testing regulations
- UK **PSTI (IoT**)
- HIPAA Evaluation Standard
- FINRA

Penetration Testing in Real Life



The need of AI in Offensive Cyber Security

- Increasing complexity of IT assets and volume of cyber threats (Mobile & Web Applications, Internal and External, Developed or Acquired apps, IoT, BYOD)
- Limitations of manual penetration testing (Cost, Time, Coverage, and Accuracy)
- Compliance and Assurance (continuous assessment to inform adaptive defences)

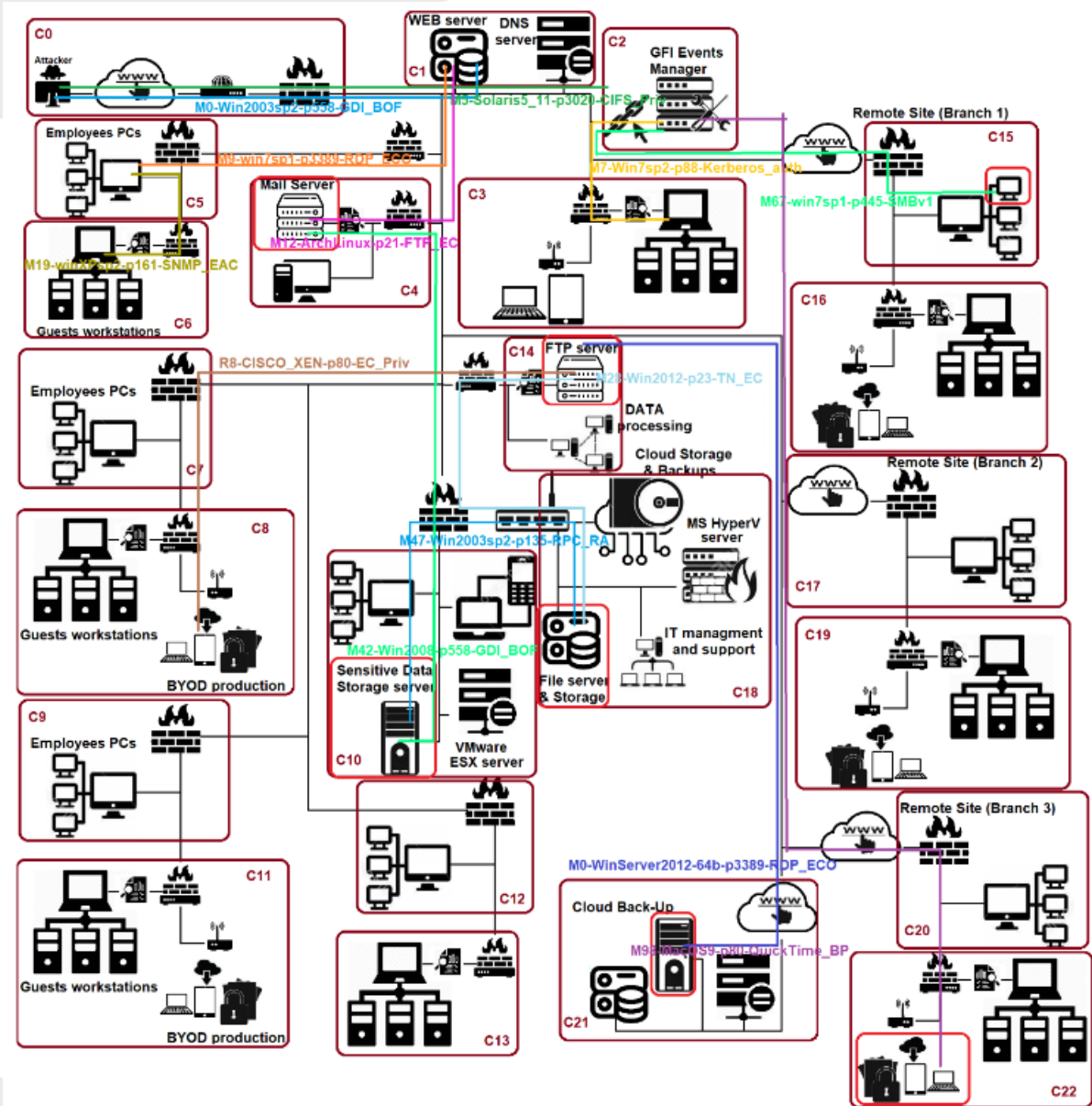


Machine Learning

- Supervised Learning: Random Forests, SVMs, and DNN achieve over 95% detection precision for known flaws.
- Feature sets include API call sequences, control-flow graphs, and syntactic code metrics to reveal buffer overflows and injection points.
- Unsupervised: Autoencoder-based anomaly detection discovers novel vulnerabilities by flagging high reconstruction errors in code.

Attack Graph Generation & Analysis

- ML-led frameworks learn to generate attack graphs from network topologies, capturing multi-step exploit paths automatically.
- Graph Neural Networks (GNNs) estimate path probabilities and impact scores.
- Automated threat modeling scales across large infrastructures, identifying complex interdependencies and potential pivot points.



(Ghanem , 2020)

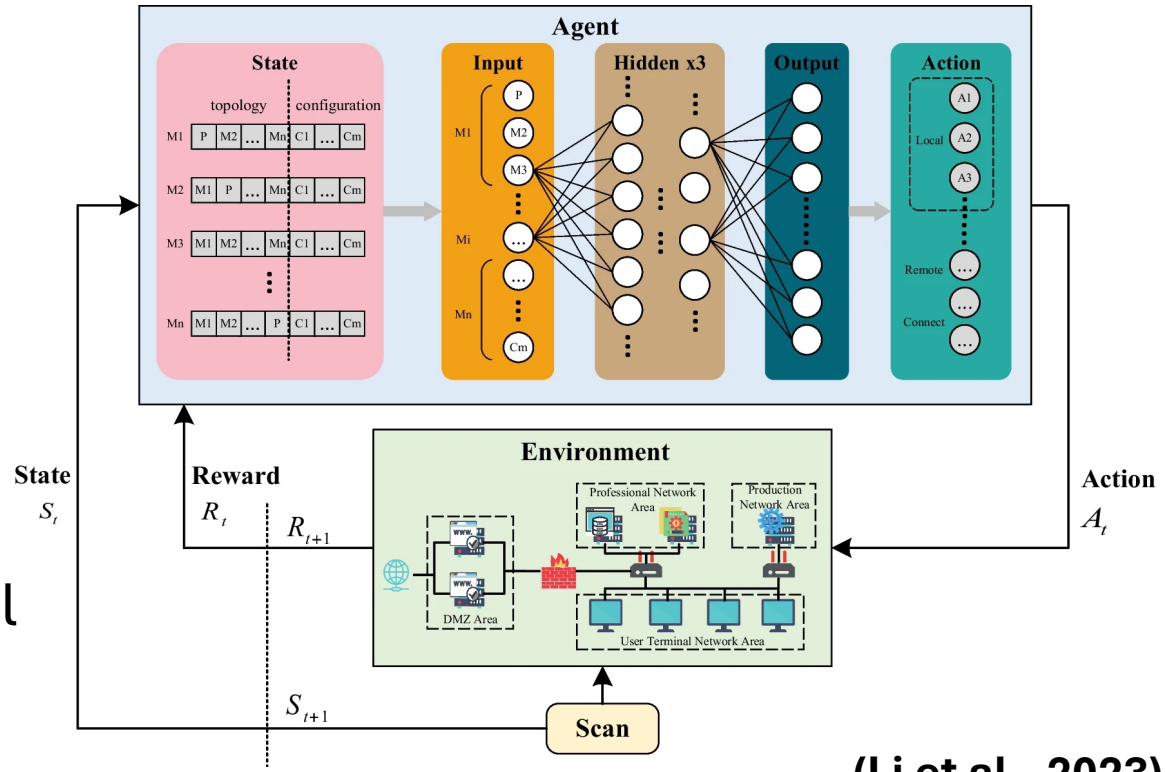
Vulnerability Prioritisation & Triage

- ML-enhanced risk scoring combines NIST CVSS with real-time threat intelligence and exploit likelihood to rank CVEs effectively.
- Exploit Prediction Scoring System (EPSS) augments statistical models with ML features to forecast exploit risks and optimize patch schedules.
- NLP pipelines using transformer architectures parse vendor advisories, mapping CVEs to playbooks and generating remediation plans.

Severity Rating	CVSS 3.1 Score	Description
CRITICAL	9.0 - 10	Exploitation of the vulnerability allows an attacker administrative-level access to systems and/or high-level data that would catastrophically impact the organization. Vulnerabilities marked CRITICAL require immediate attention and must be fixed without delay, especially if they occur in a production environment.
HIGH	7.0 – 8.9	Exploitation of the vulnerability makes it possible to access high-value data. However, there are certain pre-requisites that need to be met for the attack to be successful. These vulnerabilities should be reviewed and remedied wherever possible.
MEDIUM	4.0 – 6.9	Exploitation of the vulnerability might depend on external factors or other conditions that are difficult to achieve, like requiring user privileges for a successful exploitation. These are moderate security issues that require some effort to successfully impact the environment.
LOW	0.1 – 3.9	Vulnerabilities in the low range typically have very little impact on an organization's business. Exploitation of such vulnerabilities usually requires local or physical system access and depends on conditions that are very difficult to achieve practically.
INFORMA-TIONAL	0.0	These vulnerabilities represent significantly less risk and are informational in nature. These items can be remediated to increase security.

Deep Learning for Cloud & Web Misconfiguration Detection

- Sequence models analyse infrastructure-as-code templates and runtime logs to detect and network misconfigurations.
- Web scanners integrated with CNNs and RNNs learn from payloads, reducing false positives.
- DRL to address vulnerability analysis with actual cyber risk and thus improving misconfiguration accuracy.



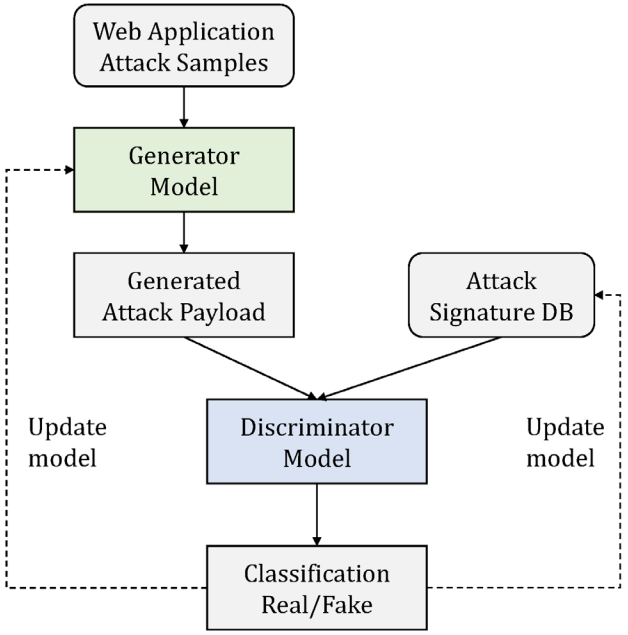
(Li et al., 2023)

Deep Learning Challenges

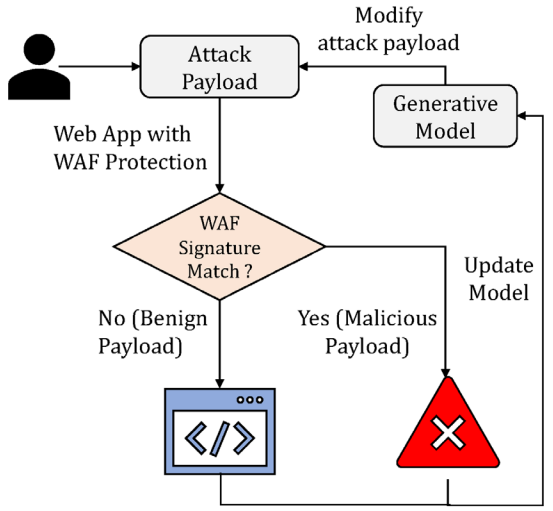
- Data Scarcity & Labeling Burden.
- Adversarial Evasion Risks.
- Lack of Interpretability.
- Scalability & Resource Constraints.

Generative Adversarial Network (GAN)

- GANs generate polymorphic payload variants to evade signature-based defences,
- Conditional GANs focus on specific exploit classes (e.g., XSS, SQLi), enabling targeted synthesis of novel test vectors.
- When integrated into CI/CD pipelines, GAN modules perform continuous vulnerability probing in DevSecOps workflows.



(a) GAN Implementation for Web Application Pentest

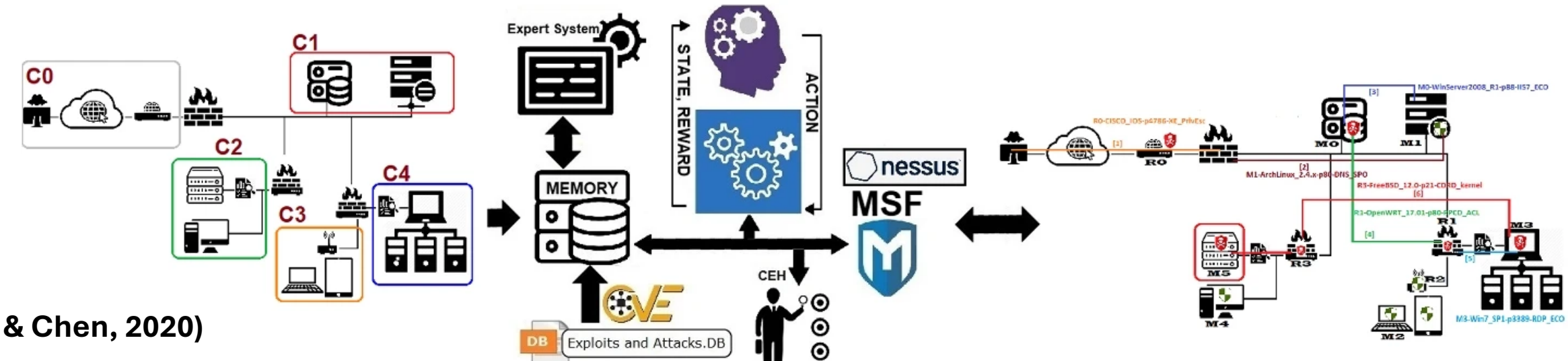


(b) GAN Data Flow for Web Application protected by Web Application Firewall (WAF)

(Chowdhary et al., 2023)

Reinforcement Learning

- Model-free RL agents traverse simulated network topologies, learning optimal strategies for privilege escalation and lateral movement.
- Reward functions balance stealth, impact, and detection avoidance, guiding multi-stage attack optimization.
- Context-aware recommendation (attack vectors) – Policy Graphs



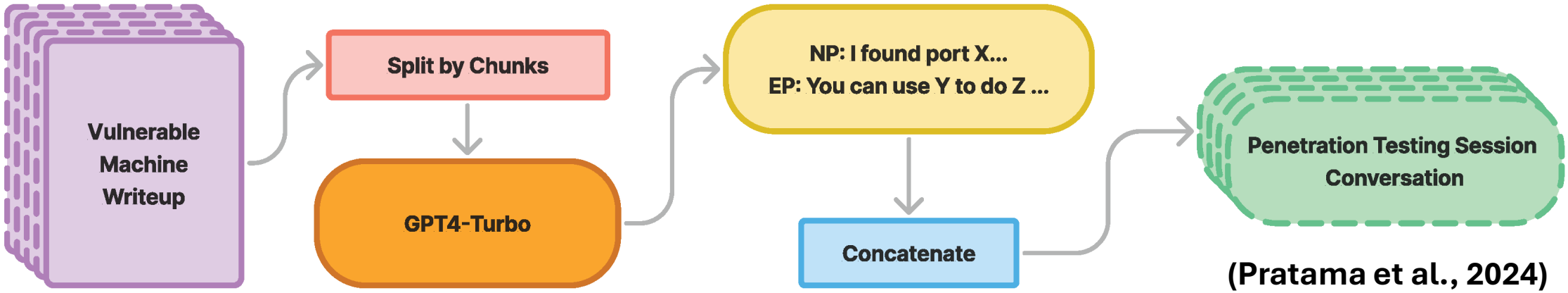
(Ghanem & Chen, 2020)

Reinforcement Learning Challenges

- High Sample Complexity & Environment Modelling
- Sparse & Shaping Rewards
- Safety & Unintended Actions
- Transferability & Domain Shift

Large Language Models

- GPT-style LLMs can draft reconnaissance scripts, exploit code snippets, and attack plans from natural language prompts.
- Adversarial prompt engineering uncovers injection vulnerabilities in AI-powered pentesting assistants.
- Case studies show LLMs generating Metasploit modules and custom payloads with minimal human oversight.



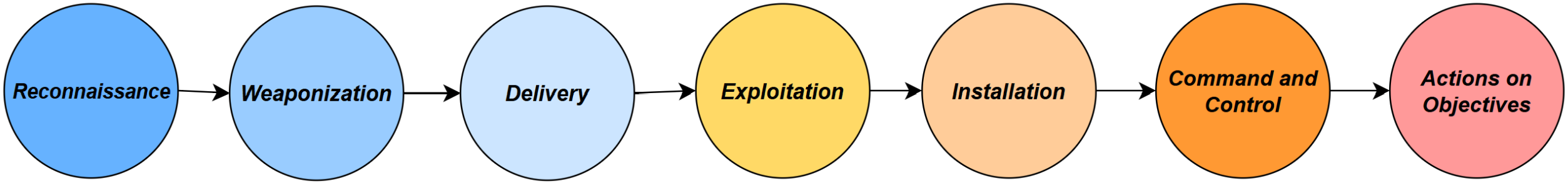
(Pratama et al., 2024)

Large Language Models Challenges

- Hallucinations & Inaccurate Exploit Suggestions
- Prompt Injection & Adversarial Manipulation
- Domain Adaptation Limitations- LLMs often lack a nuanced understanding of specific network architectures or web frameworks, requiring extensive fine-tuning.
- Ethical & Legal Misuse Potential

The parallel PT and MTD - the Cyber Kill Chain

- Modern Penetration Testing aims to inform Moving Target Defence (MTD) decisions
- Cyber Kill Chain is the way PT and MTD meet.
- Understanding attack progression helps in applying PT (offensive) MTD (defensive) cyber security.



Case Study: PT as Partially Observable Markov Decision Process

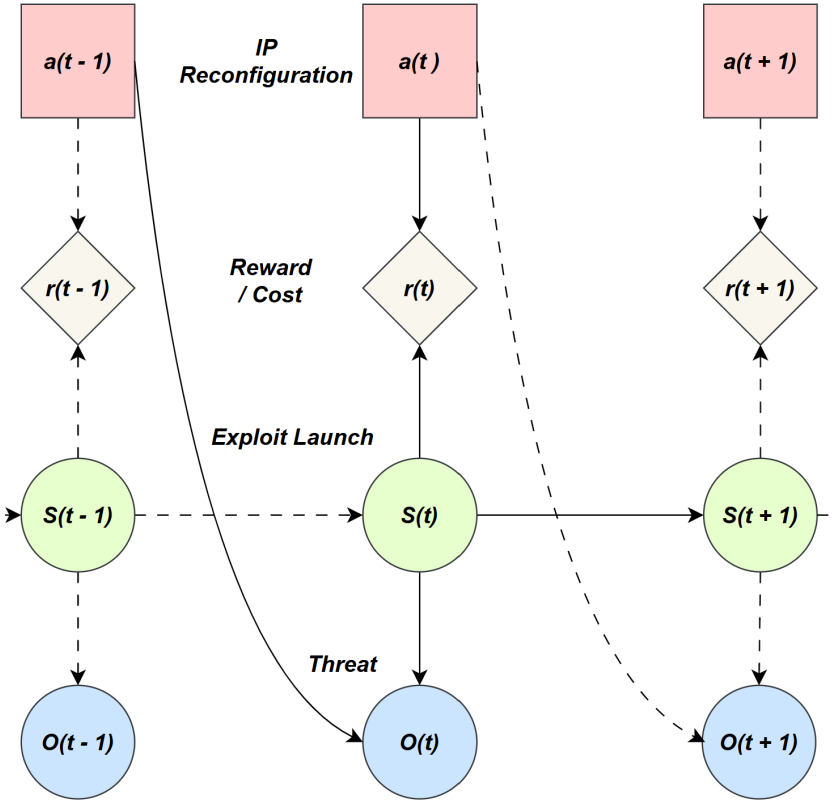
Real-world cybersecurity environments often have incomplete information

Key Components:

- **States (S):** System conditions (e.g., secure, under attack, compromised).
- **Actions (A):** Defender choices (e.g., apply MTD, do nothing)
- **Transition Probabilities (T(s' | s, a)):** Likelihood of moving between states after an action.
- **Observations (O):** Limited signals instead of knowing the exact attack state.
- **Observation Model ($\mathbf{Z}(\mathbf{o} \mid \mathbf{a}, \mathbf{s}')$):** Links hidden states to observable evidence.
- **Rewards (R):** Immediate Reward & Operational Cost.

Challenges:

- **Curse of Dimensionality:** Exponentially growth of possible states.
- **Curse of History:** Increasing memory and processing demands over time.



Network Penetration Testing (POMDP)

1. State Definitions

- ❑ Network Posture: Open ports, active services, patch levels, firewall rules, authentication mechanisms.
- ❑ Vulnerability Status: Known CVEs, misconfigurations (e.g., default credentials, unencrypted protocols).
- ❑ Access Level: Current privileges (e.g., unauthenticated, user, admin, root).
- ❑ Defender Awareness: Detection alerts triggered, honeypots activated, defensive countermeasures (e.g., IP blocking)

2. Actions

- ❑ Reconnaissance: Port scanning, service fingerprinting, vulnerability scanning.
- ❑ Exploitation: Deploy payloads (e.g., SQLi, buffer overflow), phishing attempts, credential brute-forcing.
- ❑ Persistence: Install backdoors, escalate privileges, exfiltrate data.
- ❑ Evasion: Clear logs, spoof IPs, use encryption to avoid detection.

3. Observation

- ❑OS Detected, OS-Undetected
- ❑Exploit-Overtime

4. Rewards

- ❑Positive: Successful exploit (+20), privilege escalation (+30), data exfiltration (+50).
- ❑Negative: Triggered alerts (-15), blocked IP (-25), system crash (-40), failed exploit (-10).

In Practice

```
States:

Internet
....
M7
M7-win7sp1
M7-win7sp1-p445
M7-win7sp1-p445-SMBv1
M7-win7sp1-p445-SMBv1-vulnerable-CVE-2017-0272
M7-win7sp1-p445-SMBv1-compromised-CVE-2017-0272
M7-win7sp1-p445-SMBv1-secured-CVE-2017-0272
M7-win7sp1-p445-SMBv1-vulnerable-CVE-2017-0277
M7-win7sp1-p445-SMBv1-compromised-CVE-2017-0277
M7-win7sp1-p445-SMBv1-secured-CVE-2017-0277
M7-win7sp1-p445-SMBv1-vulnerable-CVE-2017-0278
M7-win7sp1-p445-SMBv1-compromised-CVE-2017-0278
M7-win7sp1-p445-SMBv1-secured-CVE-2017-0278
M7-win7sp1-p3389
M7-win7sp1-p3389-RDP_ECO
M7-win7sp1-p3389-RDP_ECO-vulnerable-CVE-2018-8494
M7-win7sp1-p3389-RDP_ECO-compromised-CVE-2018-8494
M7-win7sp1-p3389-RDP_ECO-secured-CVE-2018-8494
M7-win7sp1-p3389-RDP_ECO-vulnerable-CVE-2018-8550
M7-win7sp1-p3389-RDP_ECO-compromised-CVE-2018-8550
M7-win7sp1-p3389-RDP_ECO-secured-CVE-2018-8550
M7-win7sp1-p3389-RDP_ECO-vulnerable-CVE-2017-11885
M7-win7sp1-p3389-RDP_ECO-compromised-CVE-2017-11885
M7-win7sp1-p3389-RDP_ECO-secured-CVE-2017-11885
M7-win7sp1-p3389-RDP_ECO-vulnerable-CVE-2016-7260
M7-win7sp1-p3389-RDP_ECO-compromised-CVE-2016-7260
M7-win7sp1-p3389-RDP_ECO-secured-CVE-2016-7260
M7-win7sp1-p88

R3
R3-CISCO_XEN
R3-CISCO_XEN-p68
R3-CISCO_XEN-p68-DHCP_Dos
R3-CISCO_XEN-p68-DHCP_Dos-vulnerable-CVE-2019-1814
R3-CISCO_XEN-p68-DHCP_Dos-compromised-CVE-2019-1814
R3-CISCO_XEN-p68-DHCP_Dos-secured-CVE-2019-1814
R3-CISCO_XEN-p68-DHCP_Dos-vulnerable-CVE-2017-3864
R3-CISCO_XEN-p68-DHCP_Dos-compromised-CVE-2017-3864
R3-CISCO_XEN-p68-DHCP_Dos-secured-CVE-2017-3864
R3-CISCO_XEN-p68-DHCP_Dos-vulnerable-CVE-2015-0578
R3-CISCO_XEN-p68-DHCP_Dos-compromised-CVE-2015-0578
R3-CISCO_XEN-p68-DHCP_Dos-secured-CVE-2015-0578
R3-CISCO_XEN-p80
R3-CISCO_XEN-p80-EC_Priv
R3-CISCO_XEN-p80-EC_Priv-vulnerable-CVE-2018-0437
R3-CISCO_XEN-p80-EC_Priv-compromised-CVE-2018-0437
R3-CISCO_XEN-p80-EC_Priv-secured-CVE-2018-0437
R3-CISCO_XEN-p80-EC_Priv-vulnerable-CVE-2016-6473
R3-CISCO_XEN-p80-EC_Priv-compromised-CVE-2016-6473
R3-CISCO_XEN-p80-EC_Priv-secured-CVE-2016-6473
R3-CISCO_XEN-p80-EC_Priv-vulnerable-CVE-2016-0705
R3-CISCO_XEN-p80-EC_Priv-compromised-CVE-2016-0705
R3-CISCO_XEN-p80-EC_Priv-secured-CVE-2016-0705
R3-CISCO_XEN-p80-EC_Priv-vulnerable-CVE-2013-1100
```

```
Actions:

Initiate
MachineStatus
OSDetect
OSCheck
PortProbv1
PortProbv2
PortProbv3
PingSweep
TraceRoute
SVCDetect
SVCCheck
VulAssess
Exploit
Re-Exploit
Pivot
ShellPersist
PrivEscalation
Terminate
Give_Up
```

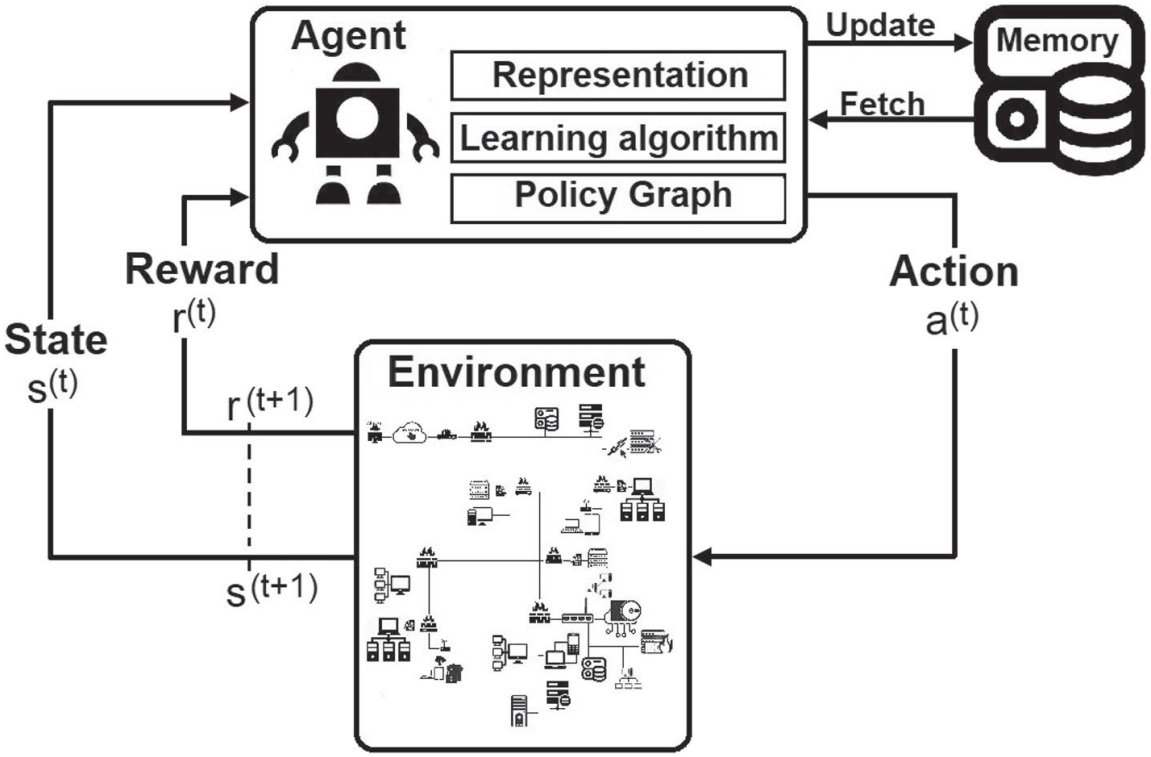
```
Observations:
.....
M5-Off
M5-On
M5-OSDetectedWinServer2012
M5-OSUndetected
M5-PortDetected-p23
M5-PortDetected-p135
M5-PortDetected-p558
M5-PortUnDetected
M5-SVCDetected-TN_EC
M5-SVCDetected-RPC_RA
M5-SVCDetected-GDI_BOF
M5-SVCUnknown
M5-VulAss-TN_EC
M5-VulAss-RPC_RA
M5-VulAss-GDI_BOF
M5-VulAssNone
M5-Exploited-TN_EC
M5-Exploited-RPC_RA
M5-Exploited-GDI_BOF
M5-Secure-TN_EC
M5-Secure-RPC_RA
M5-Secure-GDI_BOF
.....
.....
Internet-M5-Pivot
M5-Internet-Pivot
M1-M5-Pivot
M5-M1-Pivot
M2-M5-Pivot
M3-M5-Pivot
M5-M0-Pivot
M5-M3-Pivot
M4-M5-Pivot
M5-M4-Pivot
M5-Terminal-Pivot
Terminal-M5-Pivot
M5-Escal-User
M5-Escal-Root
....
Test-Acheived
Test-Partially
Test-Stopped
Test-Overtime
```

```
R : Exploit : M5-Solaris5_11-p3260-iSCSI-vulnerable : M5-Solaris5_11-p3260-iSCSI-compromised : M5-Exploited-iSCSI 1.00
R : Exploit : M5-Solaris5_11-p23-TN_Daemon-vulnerable : M5-Solaris5_11-p23-TN_Daemon-compromised : M5-Exploited-TN_Daemon 1.00
R : Exploit : M5-Solaris5_11-p3020-CIFS_Priv-vulnerable : M5-Solaris5_11-p3020-CIFS_Priv-compromised : M5-Exploited-CIFS_Priv 1.00
R : Exploit : M5-Solaris5_11-p3260-iSCSI-vulnerable : M5-Solaris5_11-p3260-iSCSI-secured : M5-Secure-iSCSI -1.00
R : Exploit : M5-Solaris5_11-p23-TN_Daemon-vulnerable : M5-Solaris5_11-p23-TN_Daemon-secured : M5-Secure-TN_Daemon -1.00
R : Exploit : M5-Solaris5_11-p3020-CIFS_Priv-vulnerable : M5-Solaris5_11-p3020-CIFS_Priv-secured : M5-Secure-CIFS_Priv -1.00
```

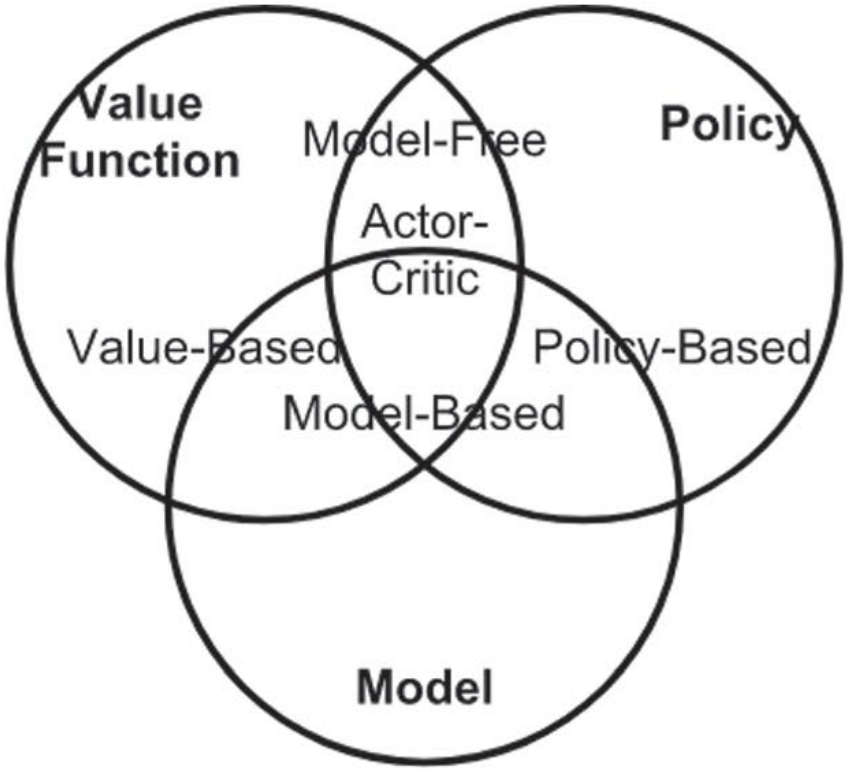
```
# Machine 12 OS detection Transition Probabilities

T: OSDetect : M12 : * 0.00
T: OSDetect : M12 : M12-WinVistaSP1 0.08
T: OSDetect : M12 : M12-WinXPSP2 0.42
T: OSDetect : M12 : M12-WinXPSP3 0.5

# Machine 12 OS detection Observation Probabilities

O: OSDetect : M12 : * 0.00
O: OSDetect : M12 : M12-OSDetectedWinXPSP3 0.49
O: OSDetect : M12 : M12-OSDetectedWinXPSP2 0.40
O: OSDetect : M12 : M12-OSDetectedWinVistaSP1 0.07
O: OSDetect : M12 : M12-OSUndetected 0.04
```

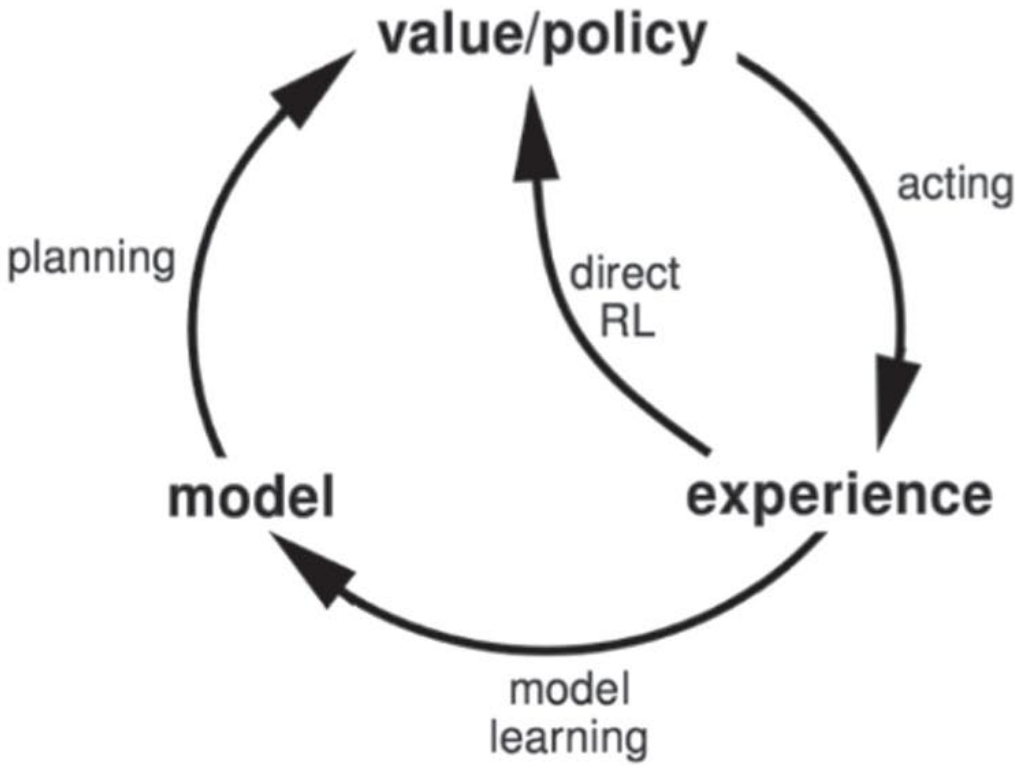
Reinforcement Learning

- **Basic Concept:** Learning through interaction with the environment
- **Key Elements:** Agent, Environment, Actions, Rewards, Policy
- **Learning Goal:** Maximise cumulative reward



Learning approaches (from RL case)





Exploration vs. Exploitation

Concepts: Balancing learning new strategies (exploration) vs. using known effective responses (exploitation)

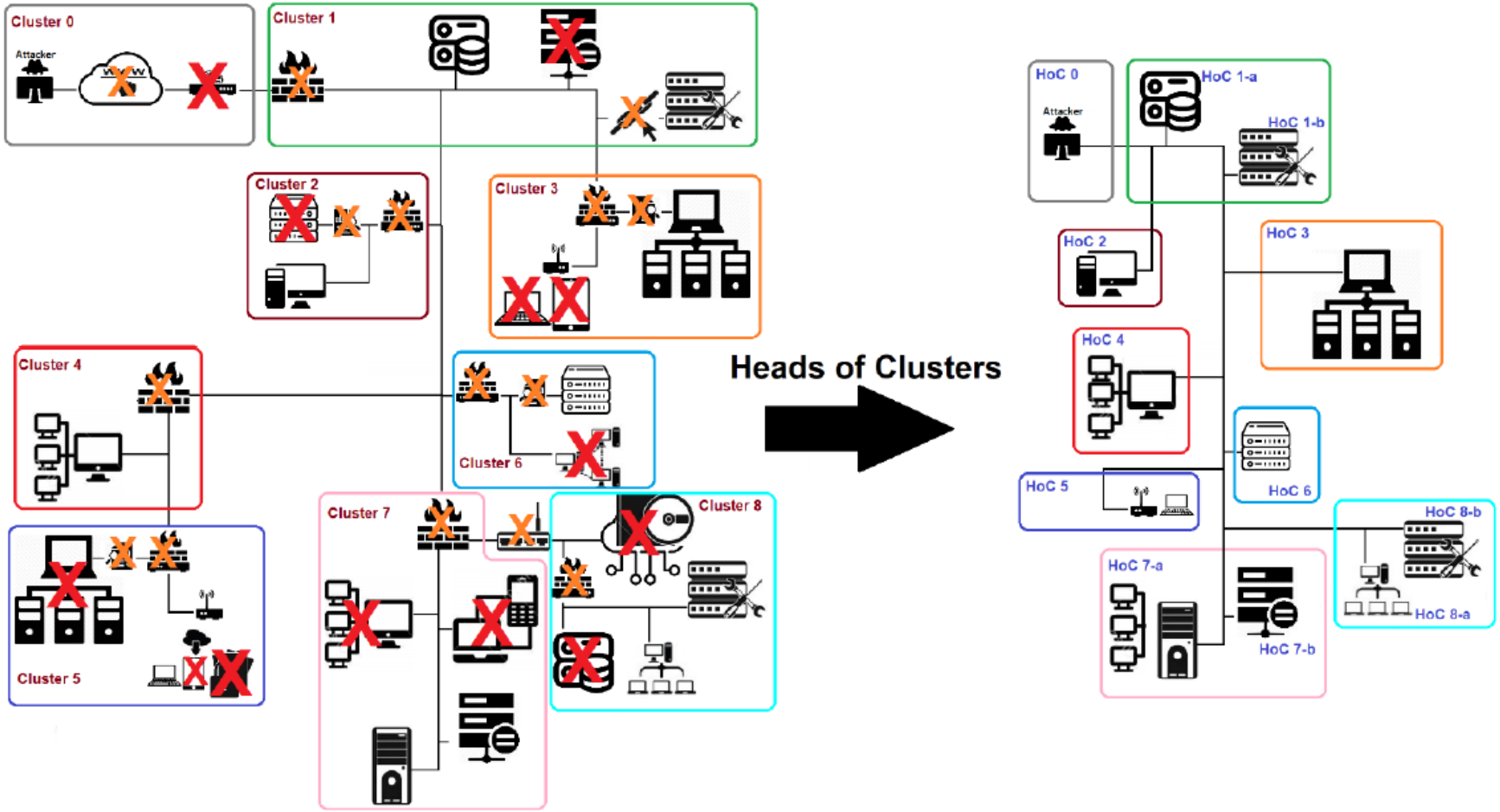
- High stakes of exploration in live environments
- Safe exploration methods that provides good exploitation

Reward Engineering and Computational cost

- Reward functions that align with security goals and constraints.
- Security actions vs operational costs to avoid disruptions.
- Sparse rewards via intrinsic motivation for rare threats.
- Reward shaping and risk-aware objectives for threats.

Domain Understanding is the key

Think like a hacker



Challenges in modelling

High Dimensionality: Large number of potential states and actions

Dynamic Threat Landscape: Evolving attack patterns

Adversarial Noise: Intentional obfuscation by attackers

Sparse and Delayed Rewards: Difficulty in timely feedback for policy updates

Cross-Cutting Infrastructure & Web App Challenges

- Integration Complexity, custom plugins and orchestration logic.
- Real-Time Adaptability.
- Adversarial Defence Arms Race.
- Regulatory & Compliance Constraints

Integrating AI with Industry Pentesting Tools and Frameworks

- Plugin architectures for Metasploit, BrupSuite and Core Impact allow AI modules to automate exploit generation, validation, and reporting.
- End-to-end platforms unify AI-driven scanning, exploitation, and reporting, enhancing workflow efficiency and consistency.
- Human-in-the-loop frameworks ensure expert oversight, ethical compliance, and contextual relevance in AI-assisted pentesting.

Future Directions & Research Challenges

Explainable AI is crucial for interpreting model decisions in vulnerability assessment, fostering trust and regulatory adherence.

Human expert role on the control of AI-led pentesting agents to prevent misuse.

DNN promise enhanced accuracy but not the transparency in attack graph analysis and vulnerability prediction.

Alignment with **emerging regulations** (EU AI Act, NIS2) will dictate governance models and auditability standards for AI-powered security tools

Summary & Key Takeaways

1. AI potential in offensive cyber security is obvious and USPs are clear
2. Understand the field and why you doing it is crucial.
3. The need to tackle domain adversarial dynamics with rich and realistic models
4. Many AI techniques are often needed (orchestration not competition).
5. Emphasis on collaboration between theory and practice; Real-World validation with live industry scenarios is key.

Thank you!

Any Question?