Maynooth University
National University of Ireland Maynooth

Hamilton Institute

Integrating Generative LLMs and Agentic Reinforcement Learning for Autonomous Cybersecurity

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

Maynooth, 01 October 2025

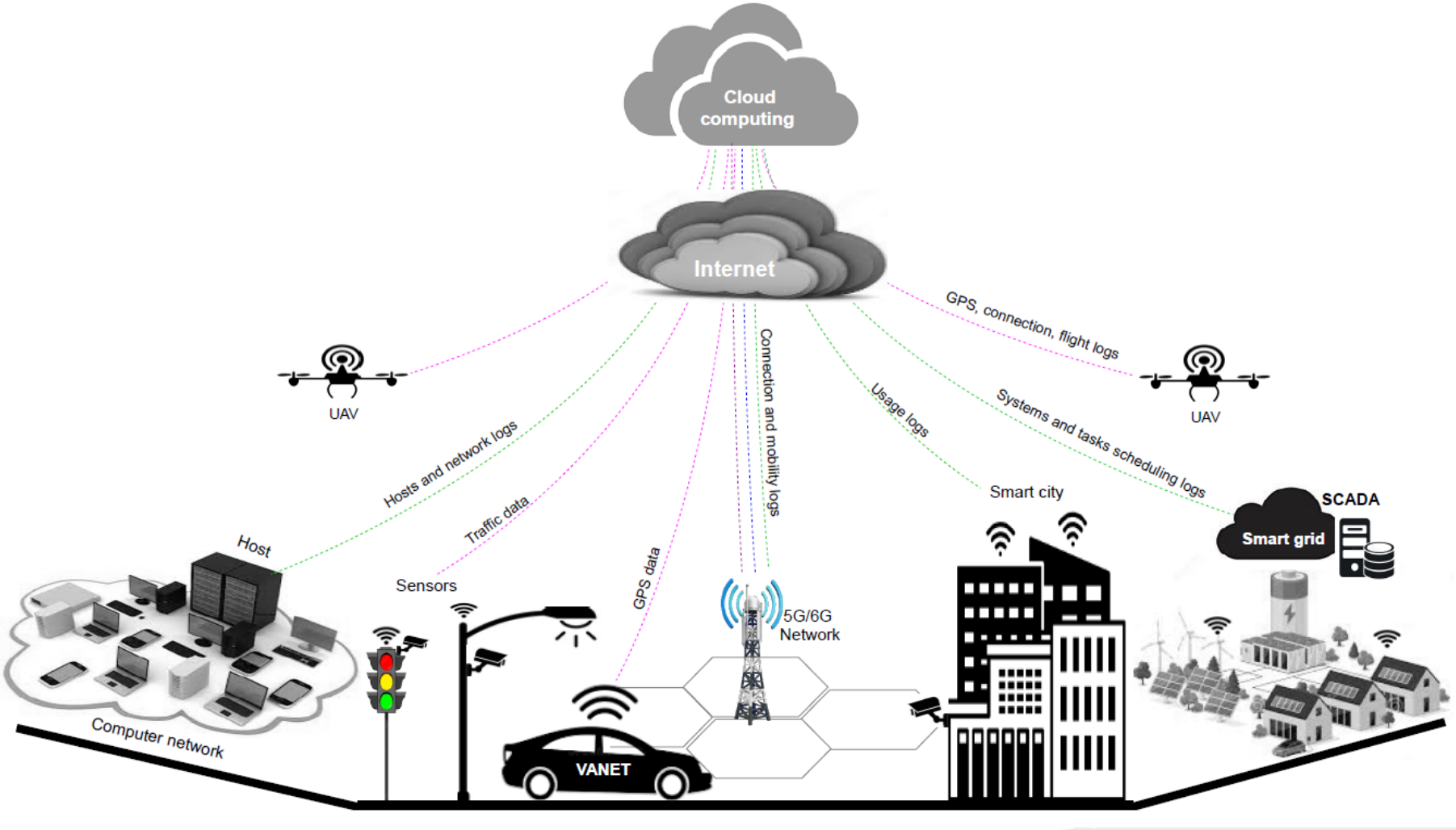
About me

- Engineer Degree in Computing (EMP), MSc in Digital Forensics (LMU), MA in Academic Practice (University of Liverpool) and PhD in Cyber Security Engineering (City, University of London).
- 15+ years in the Cyber Security Industry (Law Enforcement and Corporations)
- Certified Expert (CISSP, CPCI, Multi-GIAC Cert.)
- Currently Associate Professor, Director of the Cyber Security Research Centre (London Met) and Visiting Associate Professor at the University of Liverpool
- Senior Advisor on Cyber Resilience (CSbD) in FinTech (former Director – Resilience and Security Auditing at KROLL LLC)

Overview & Agenda

1. Introduction & Motivation
2. What Autonomous Cyber Security is about ?
3. Generative AI and LLMs
4. LLMs and RAGs in Cyber Security
5. Reinforcement Learning (RL) Fundamentals
6. MDP & POMDP Frameworks in Cybersecurity
7. Challenges in Real-World Cyber Environments
8. Bridging Theory and Practice
9. Conclusion & Future Directions

Current Cyber Security Context



Human Expert in Cyber Security?

- Very hard to define expert as it is subjective
- Often related to technical expertise and skills and validated through certifications (not about Degrees)
- Shortage of Expert and if found it is a very expensive
- Only a limited number (<5%) can afford hiring Experts

Cybersecurity in the Age of Adaptive Threats

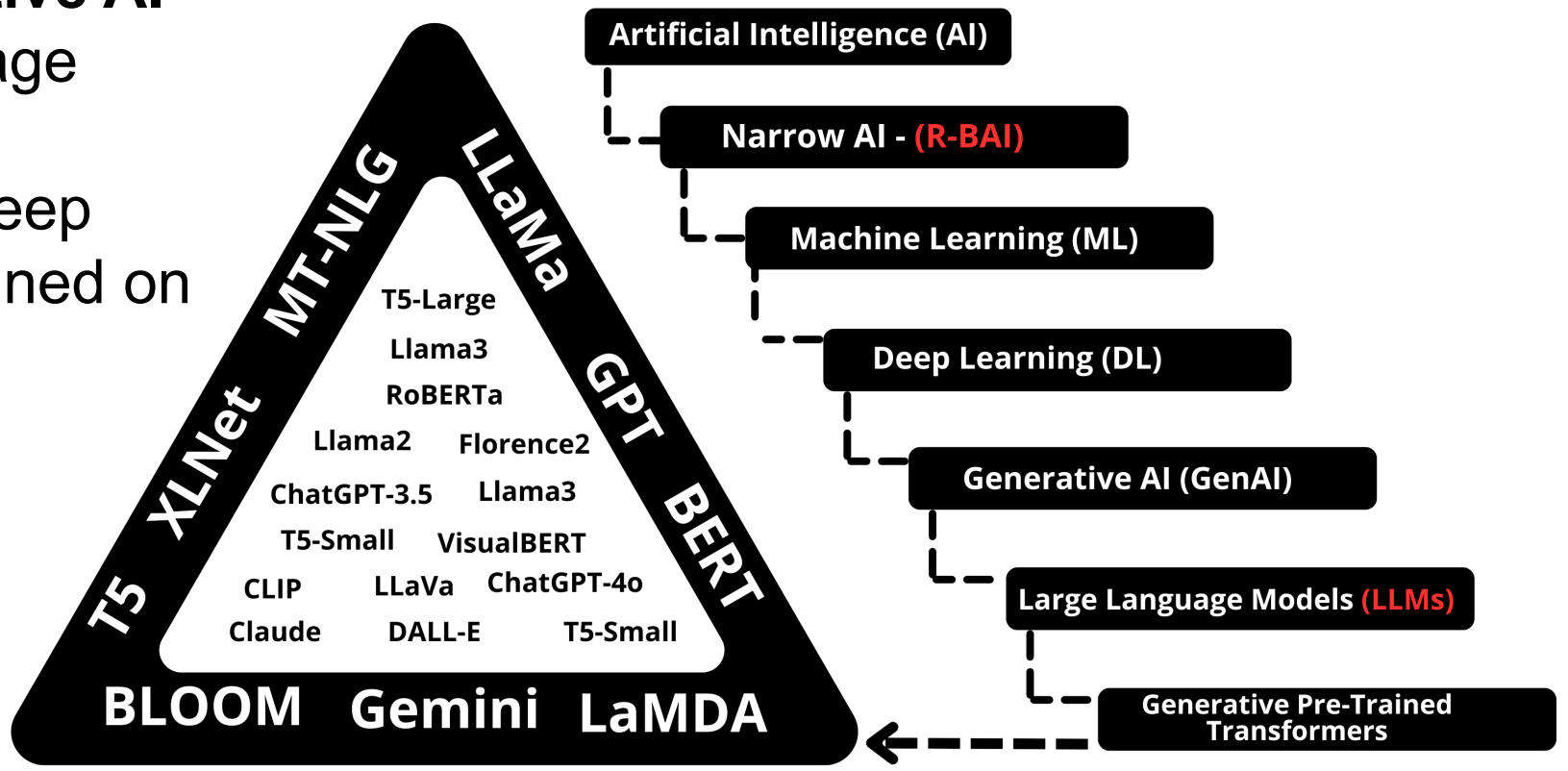
- **Static defences give attackers the upper hand** – They have time to analyse and exploit vulnerabilities.
- **Cyber threats are evolving rapidly** – APTs, zero-day attacks, and large-scale cybercrime are on the rise.
- **Traditional security lacks adaptability** – once deployed, defences remain unchanged for long periods.
- **Systems' unpredictability** – continuously changes system configurations to disrupt attackers.
- **Shifting the balance** – MTD helps defenders regain control by making attacks harder and riskier.

Autonomous Cyber Security?

- Increasing complexity and volume of cyber threats
- Limitations of manual threat detection and response
- Need for scalable, real-time, adaptive defences
- Benefits of automation: continuous monitoring, rapid response, reducing costs and minimizing errors

Large Language Models (LLMs)

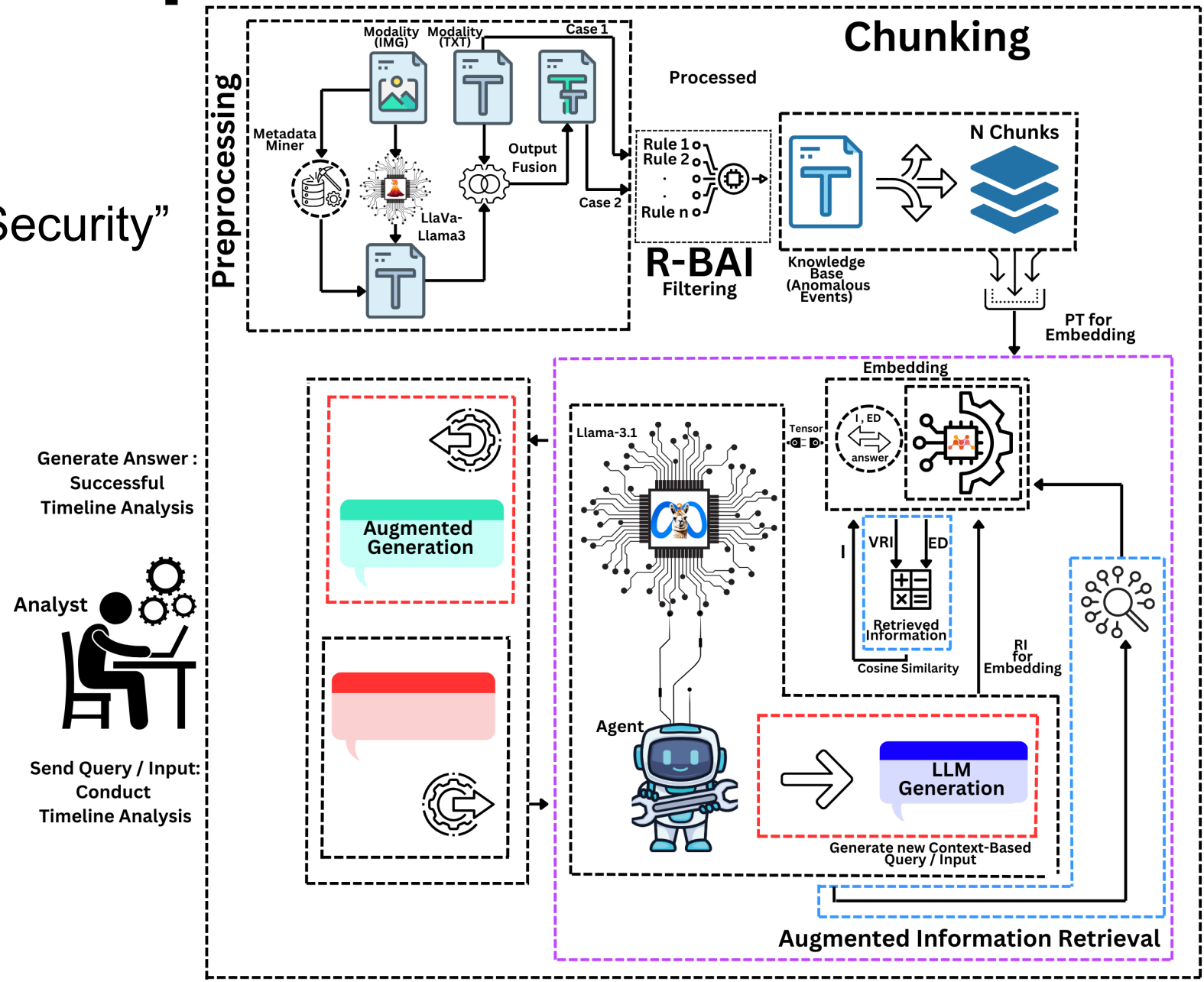
- LLMs are a branch of **Generative AI**
- Focused specifically on language processing.
- LLMs are built on advanced Deep Learning architectures and trained on massive text datasets.



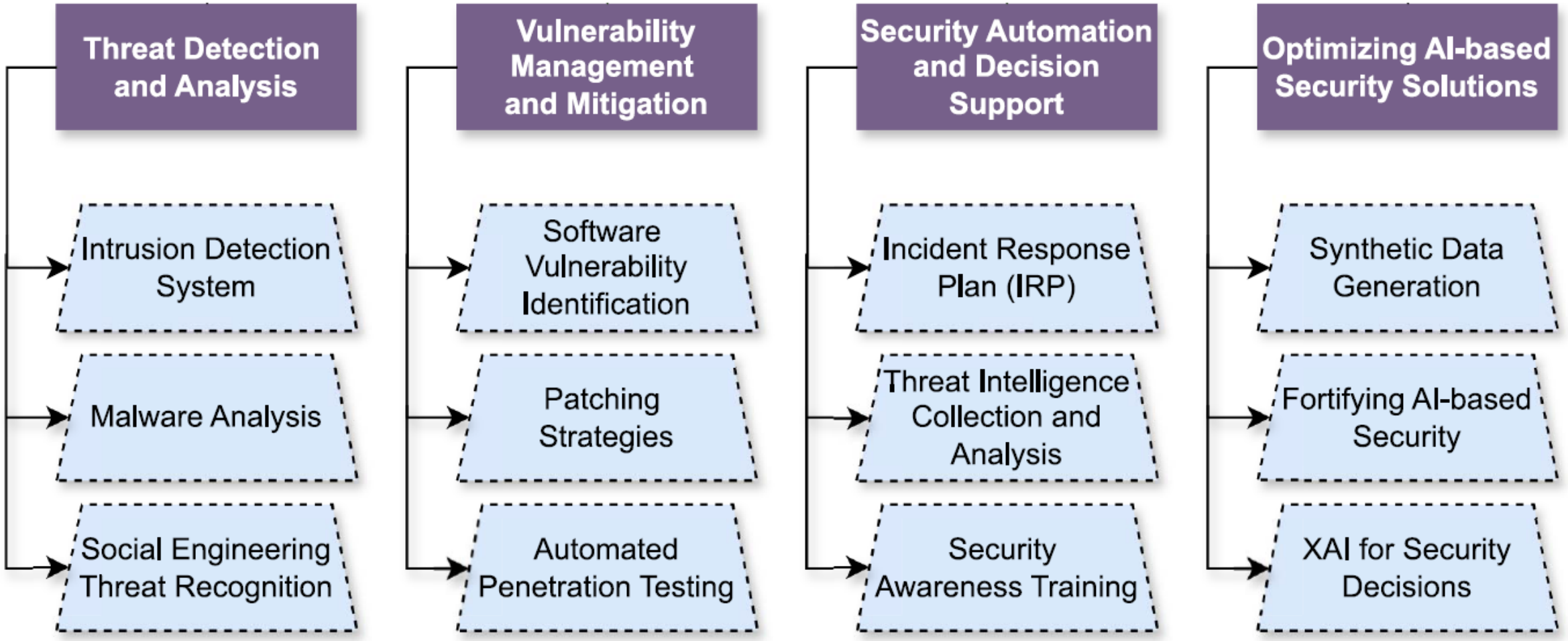
Models are trained on diverse text corpora using neural networks with billions of parameters, allowing them to capture complex language structures and subtle meanings.

LLMs in Incident Response

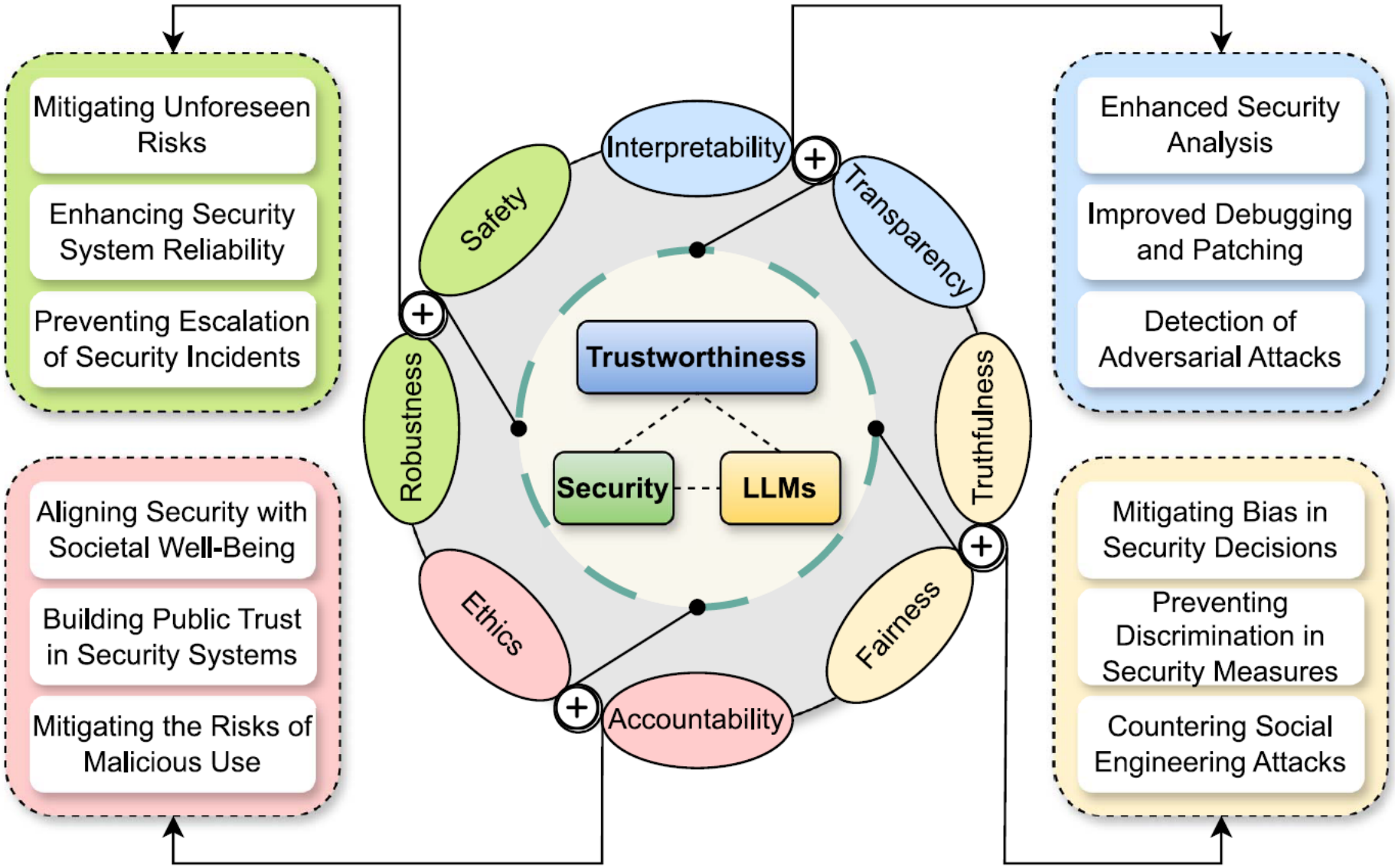
“LLMs are team player in Cyber Security”



LLMs in Cyber Security



Trustworthiness, LLMs and Security

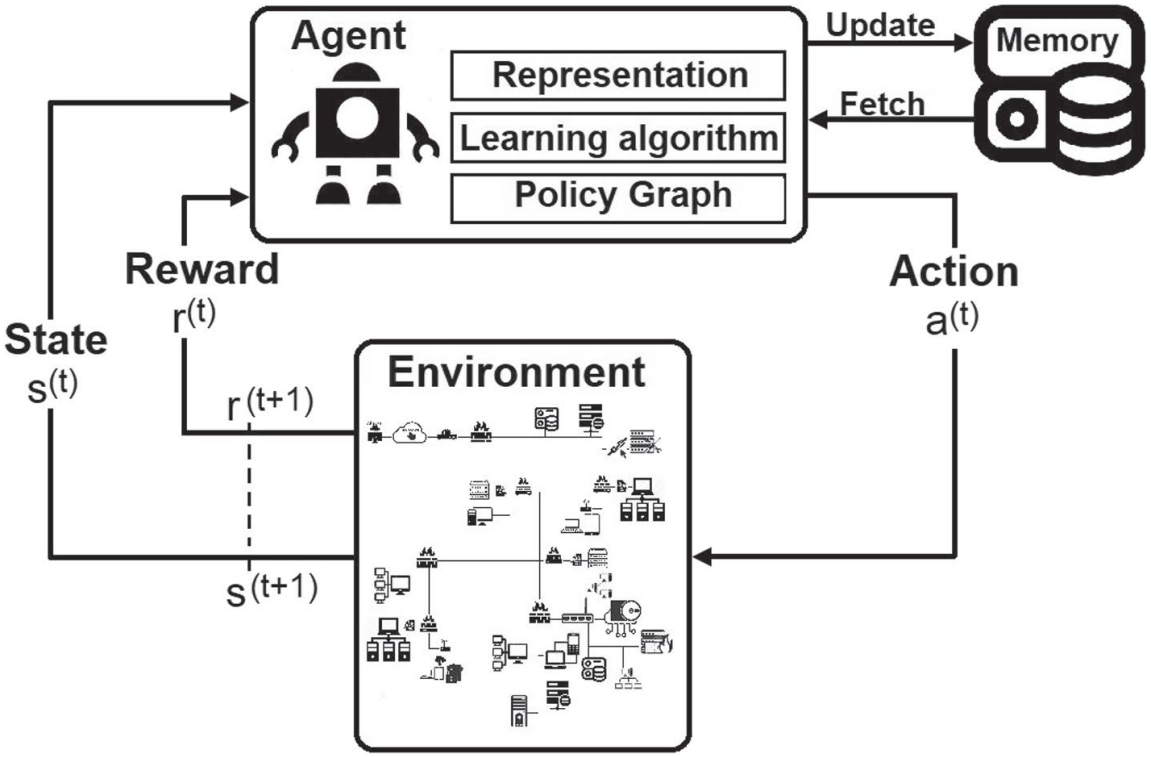


Current Research hot topic in LLMs

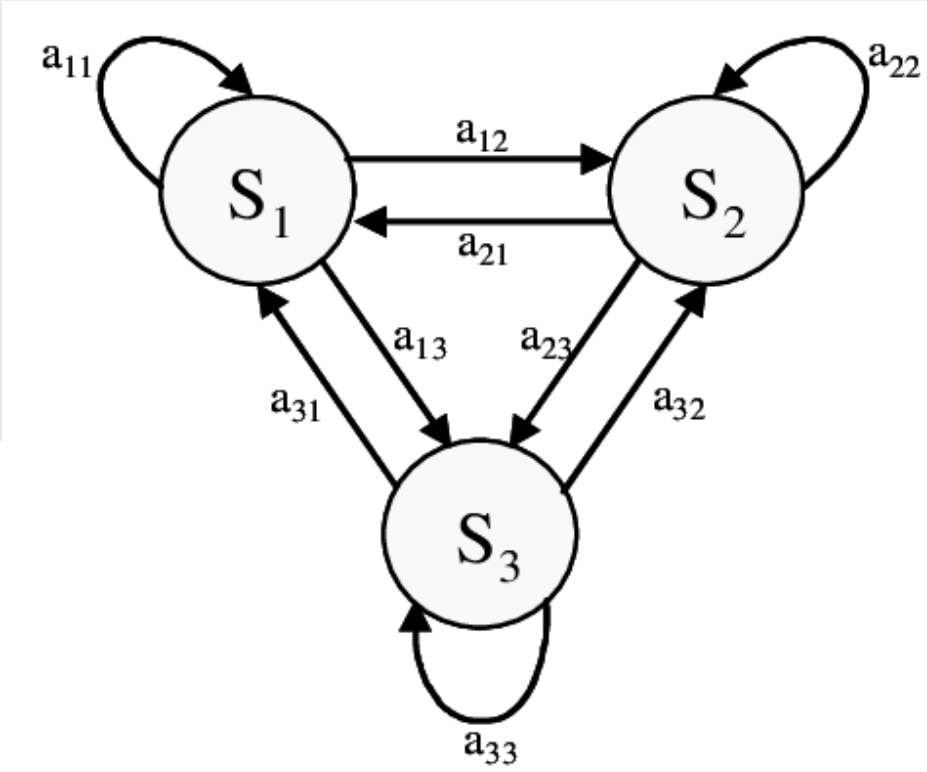
- Threat Modelling & Dataset Augmentation
- Synthesize polymorphic malware pseudocode and realistic attack narratives for red-team exercises
- Generate high-fidelity phishing exemplars and social engineering templates for training
- Create LLM-augmented datasets to improve anomaly detection and adversarial robustness

Reinforcement Learning

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



Markov Chains and MDP



Markov Chains model dynamic systems where the next state depends only on the current state, not “past history” (**Markov property**).

Absorbing Markov Chains have special 'absorbing' states where the system gets stuck permanently.

Partially Observable Markov Decision Process

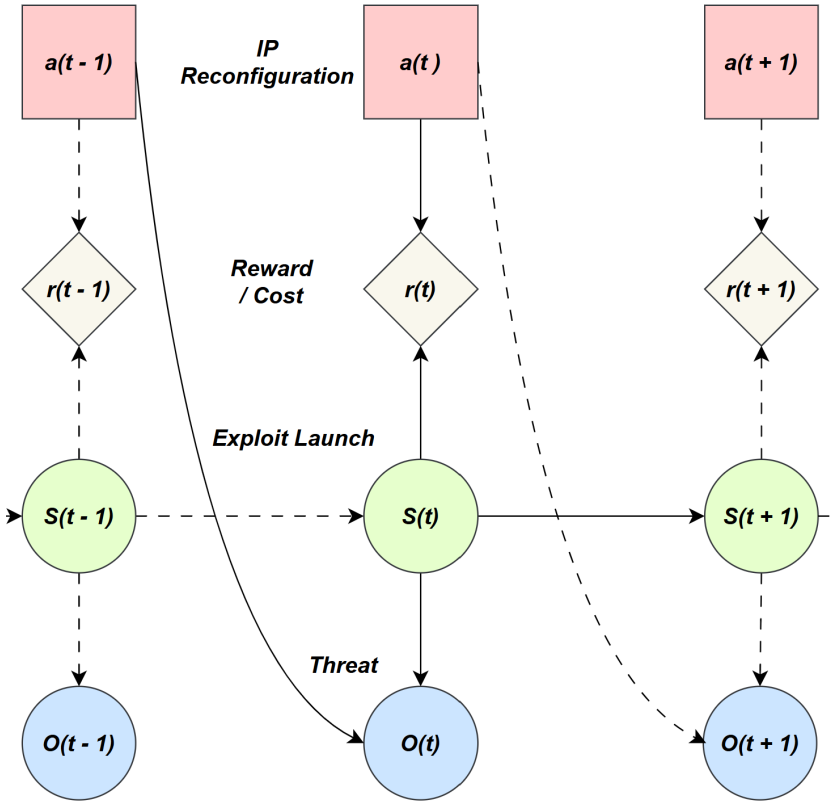
Real-world cybersecurity environments often have incomplete information

Key Components:

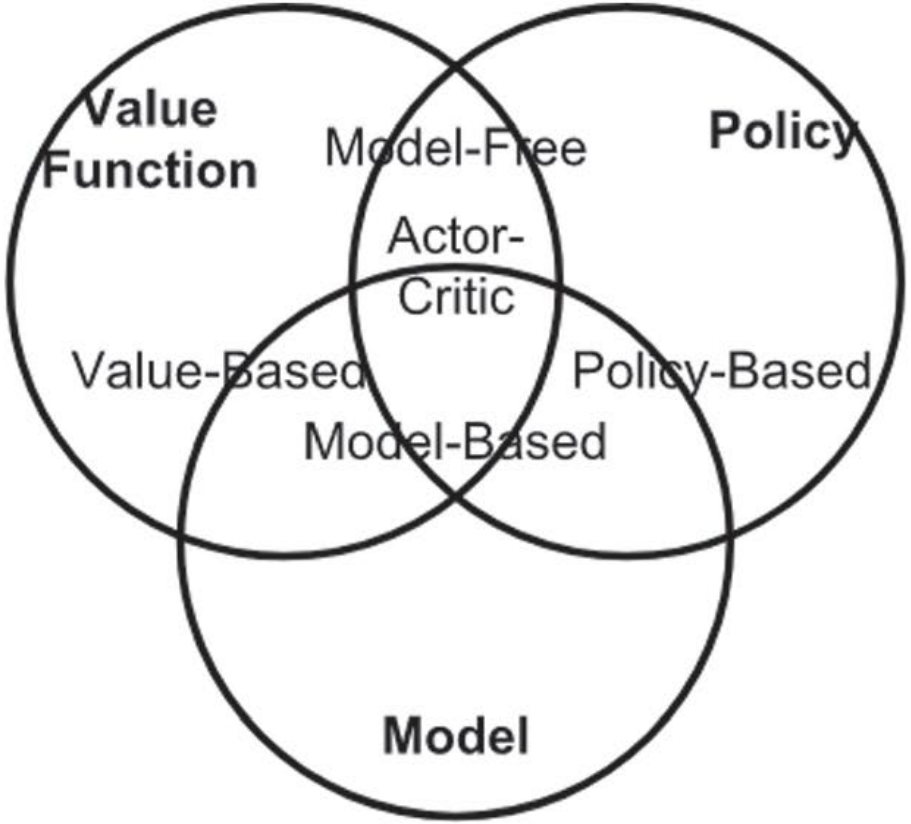
- **States (S):** System conditions (e.g., secure, under attack, compromised).
- **Actions (A):** Defender choices (e.g., apply MTD, do nothing)
- **Transition Probabilities (T(s' | s, a)):** Likelihood of moving between states after an action.
- **Observations (O):** Limited signals instead of knowing 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.

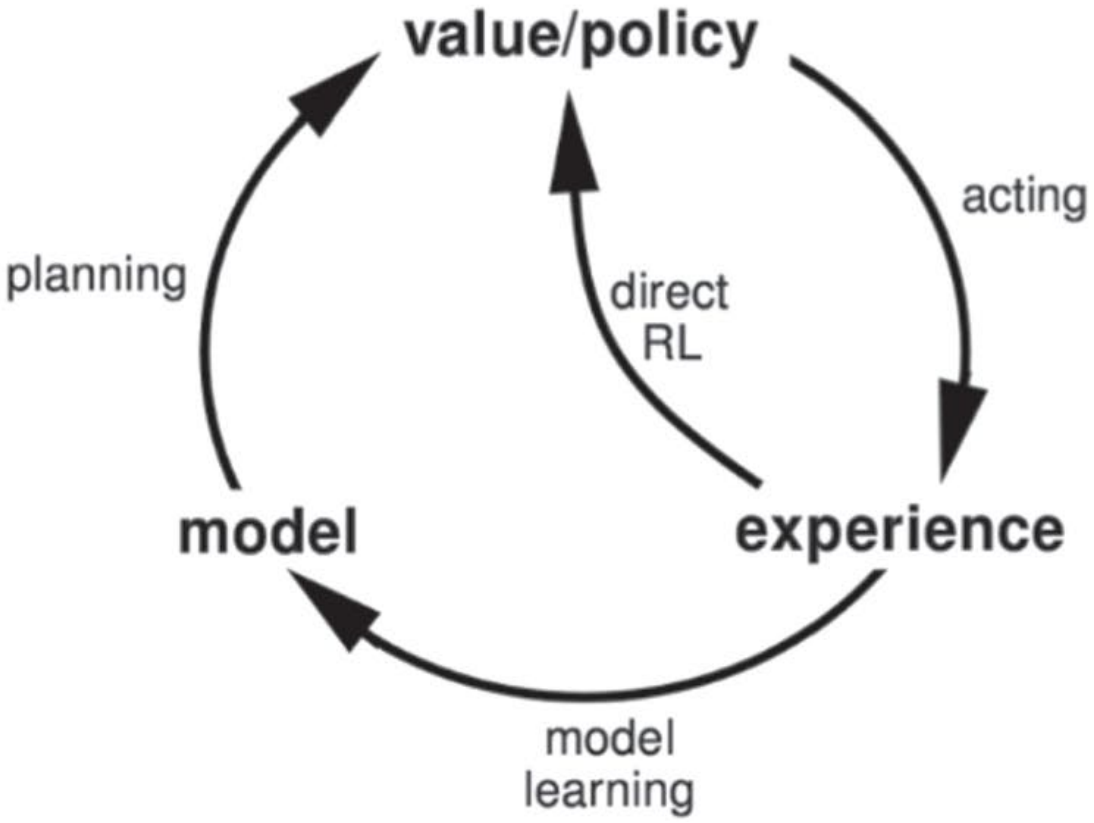
Challenges:

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



RL in Practice





Adaptive Cyber Security

Dynamic Decision-Making in Uncertainty: Continuous adaptation of the POMDP model ensures accurate decision-making.

Three-Stage Methodology:

- **Development of POMDP Model:** Captures key elements of cyber defence while incorporating uncertainty.
- **POMDP Planning Phase:** Selects and deploys MTD techniques based on updated belief states.
- **Performance Assessment:** Assesses strategy effectiveness and optimizes defensive measures.

Iterative Refinement: If objectives aren’t met, the model is recalibrated to enhance accuracy and effectiveness.

Exploration vs. Exploitation

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

Challenges:

- High stakes of exploration in live environments
- Safe exploration methods and simulations

Techniques: Model-based RL, risk-sensitive exploration

Challenges in Modelling Cyber Environments

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

Reward Engineering and Computational cost

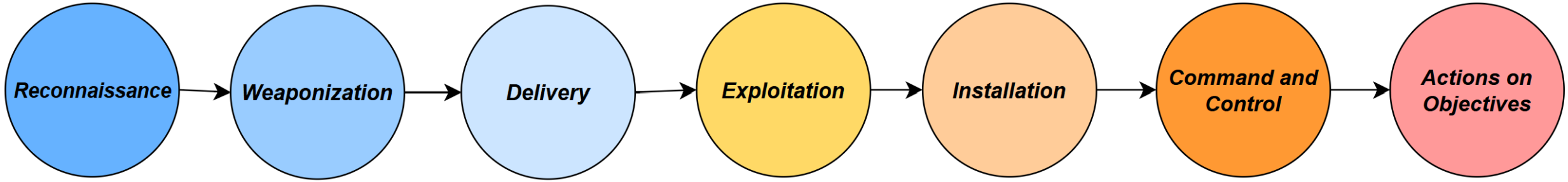
- Reward functions that align with security goals and constraints.
- Security actions vs operational costs to avoid disruptions.
- Sparse rewards via intrinsic motivation for rare threats.
- Reward shaping and risk-aware objectives for threats.

Moving Target Defence

- MTD was Introduced by **NITRD** in 2009 – Addressing cybersecurity challenges.
- **MTD rational** – Continuously changes system configurations to disrupt attacks.
- **Shell Game Analogy** – Shuffling system components increases uncertainty.
- **Dynamic Security Approach** – Increases attack complexity and costs.

MTD need to mimic Cyber Kill Chain

- Cyber Kill Chain models the structured steps of an attack.
- Understanding attack progression helps in applying MTD effectively.



Implementing MTD

MTD is promising but has challenges:

- High resource consumption and system complexity.
- Balancing security and operational manageability is difficult.
- Some MTD techniques may not be fully compatible with each other.

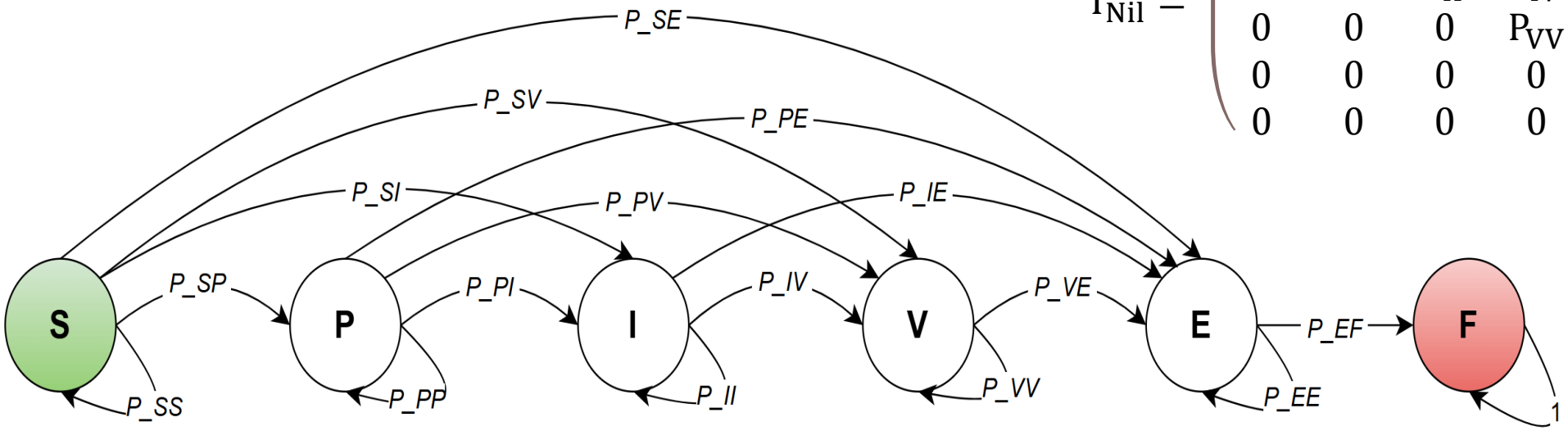
Key research challenge:

- How do we optimize MTD strategies to enhance security while ensuring availability?
- Develop a realistic POMDP to model MTD under uncertainty.
- Optimize system overhead while maximizing security effectiveness and system availability.
- Incorporate risk assessment to guide adaptive MTD decision-making.

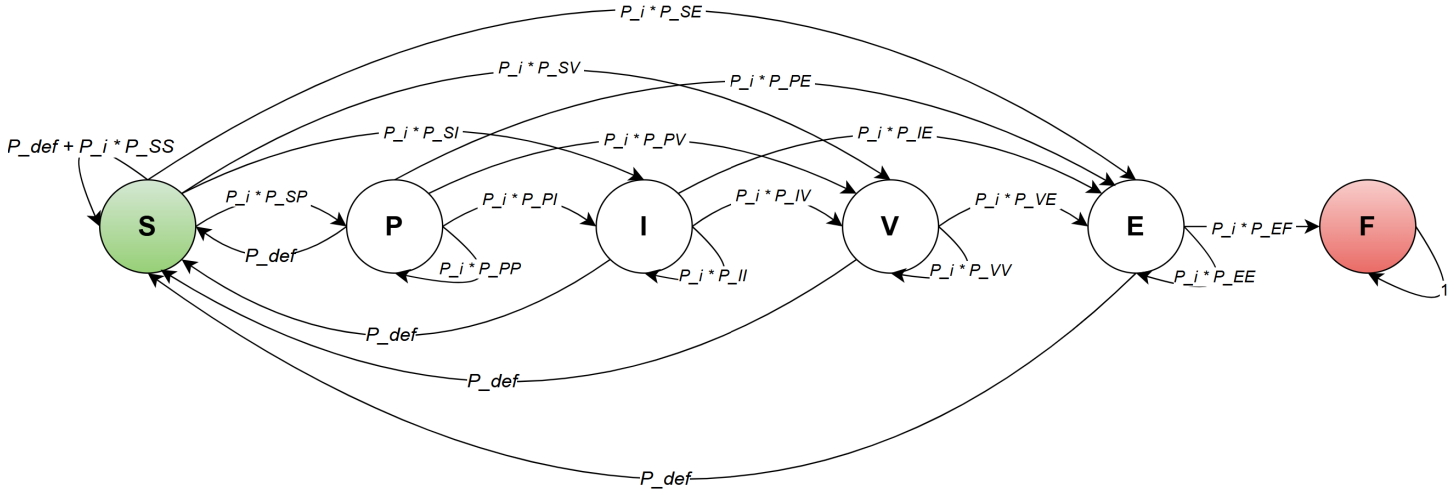
MTD Attack

- Flexible Attack Progressions: Allows step skipping.
- System can progress or stay in place but cannot move backward.
- The transitions occur under the action ’Nil’, representing an undefended system operating normally while under attack.

$$T_{Nil} = \begin{pmatrix} P_{SS} & P_{SP} & P_{SI} & P_{SV} & P_{SE} & 0 \\ 0 & P_{PP} & P_{PI} & P_{PV} & P_{PE} & 0 \\ 0 & 0 & P_{II} & P_{IV} & P_{IE} & 0 \\ 0 & 0 & 0 & P_{VV} & P_{VE} & 0 \\ 0 & 0 & 0 & 0 & P_{EE} & P_{EF} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$



MTD Defence

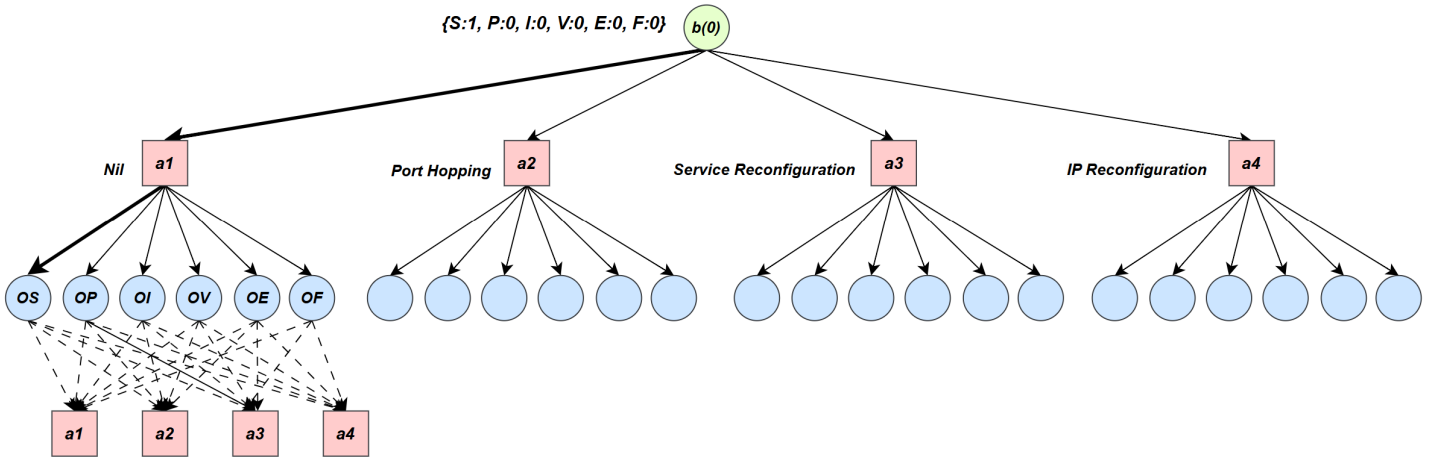


The Success Probabilities for each Defence Action:

- **Port Hopping (PH):** 0.55
- **Service Reconfiguration (SR):** 2/3
- **IP Reconfiguration (IR):** 255/256

$$T_i = \begin{pmatrix} P_{def} & 0 & 0 & 0 & 0 & 0 \\ P_{def} & 0 & 0 & 0 & 0 & 0 \\ P_{def} & 0 & 0 & 0 & 0 & 0 \\ P_{def} & 0 & 0 & 0 & 0 & 0 \\ P_{def} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & P_{def} \end{pmatrix} + (1 - P_{def}) * T_{Nil}$$

POMDP Solving



Partially Observable Monte Carlo Planning (POMCP)

- Online Method
- Uses Monte Carlo Tree Search & Particle Filtering
- Scalable for large state spaces

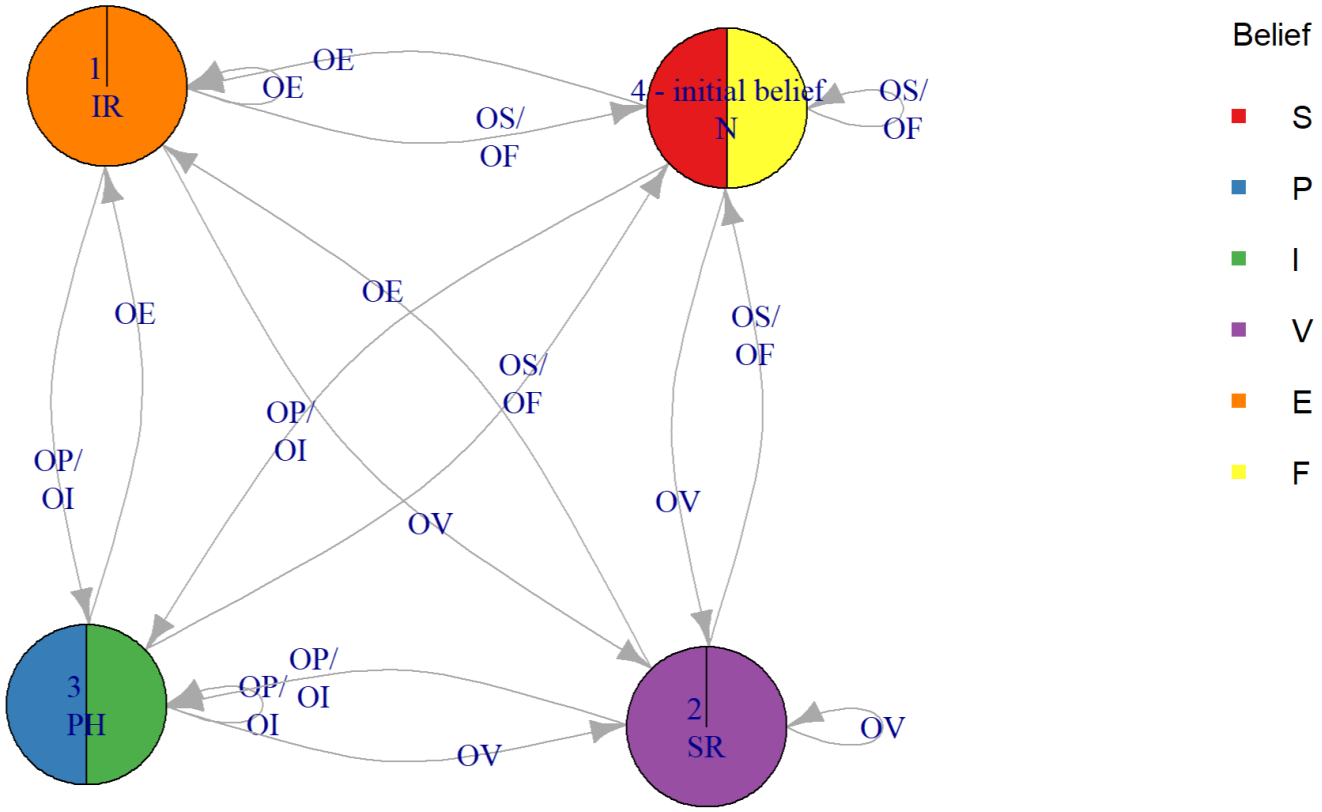
PERSEUS - Point-Based Value Iteration

- Offline Method
- Improved with Importance Sampling instead of uniform sampling
- Faster convergence & better efficiency
- Randomized Backup: Updates only a subset of belief points per iteration → reduces computation time
- Focus on High-Impact Beliefs: Prioritizes important regions of the belief space → reduces variance in updates
- **Scalable for Large POMDPs - More efficient belief selection improves policy quality**

RL and System Validation

Optimal Policy under Perfect "Detection"

- Start State -> Nil Action (N)
- Port Scan -> Port Hopping (PH)
- ICMP Scan -> Port Hopping (PH)
- Vulnerability Scan -> Service Reconfiguration (SR)
- Exploit Launch -> IP Reconfiguration (IR)
- Security Failure -> Nil Action (N)



Performance Assessment

- **Attack Success Probability (ASP):** $\boldsymbol{\psi^* = \frac{\tau_{Nil}}{\tau^*} \times 100\%}$
  - $\boldsymbol{\tau_{Nil}}$: Expected Total Number of Transitions Before Absorption where the "Nil" action is taken.
  - $\boldsymbol{\tau^*}$: Expected Total Number of Transitions Before Absorption.
- **Defence Success Probability (DSP):** $\boldsymbol{\varphi^* = \left(1 - \frac{\upsilon(s_n)}{\sum_{i=2}^{n} \upsilon(s_i)}\right) \times 100\%}$
  - $\boldsymbol{\upsilon(s_n)}$: The number of times the system reaches the final state (successful attack).
  - $\boldsymbol{\upsilon(s_i)}$: The denominator sums up the visits to all states except the first state where the system works normally.
- **System Availability:** $\boldsymbol{\vartheta^* = \frac{\zeta - \lambda^*}{\zeta} \times 100\%}$
  - $\boldsymbol{\zeta}$: The maximum network availability between decision epochs.
  - $\boldsymbol{\lambda^*}$: The expected overhead of the system.

Example- Penetration Testing (POMDP)

Mapping Network Penetration Testing to MDP/POMDP Models

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

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- Penetration Testing

```
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
```

Example- DDoS (MDP)

1. State Definitions

- **Network Status**: Traffic volume, connection rates, server load, open ports, device connectivity.
- **System Vulnerabilities**: Unpatched software, misconfigurations, weak authentication protocols.
- **Threat Signatures**: Known attack patterns (e.g., DDoS traffic signatures, malware behaviour).

2. Actions

- **Preventative Measures**: Apply patches, update firewall rules, restrict access.
- **Active Countermeasures**: Block IPs, rate limit traffic, isolate compromised nodes.
- **System Updates**: Deploy security patches, refresh threat databases.

3. Rewards

- **Positive**: Mitigated attack (e.g., stopped DDoS), minimal downtime.
- **Negative**: False positives (blocking legitimate users), false negatives (undetected breaches), resource overhead.

In Practice- DDoS Attack

States:

- S0: Normal traffic.
- S1: Low-intensity DDoS.
- S2: High-intensity DDoS.
- S3: System crash.

Actions:

- A1: Monitor/analyse traffic.
- A2: Block suspicious IPs.
- A3: Deploy traffic scrubbing.

Observations:

- Traffic spikes, geographic anomalies, server response latency.
- Uncertainty: Legitimate surge (e.g., viral event) vs. attack.

Rewards:

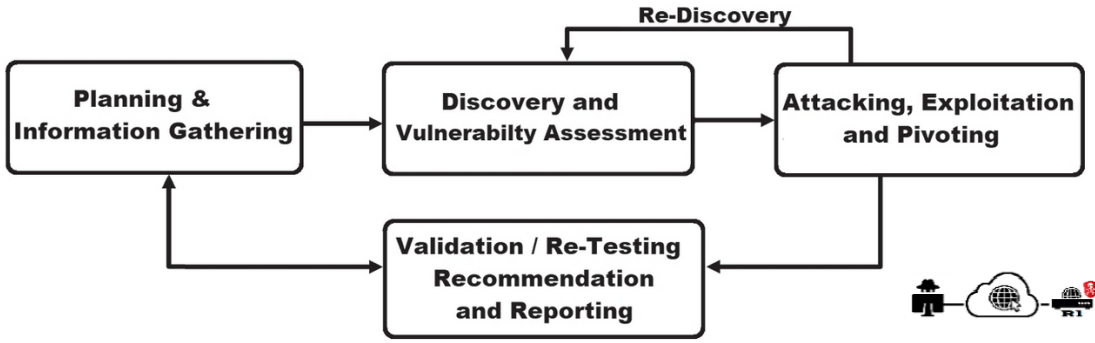
- R(S1,A2) R(S1,A2): +10 if attack stops, -5 if legitimate users blocked.
- R(S2,A3) R(S2,A3): +20 for mitigation, -2 for operational cost.

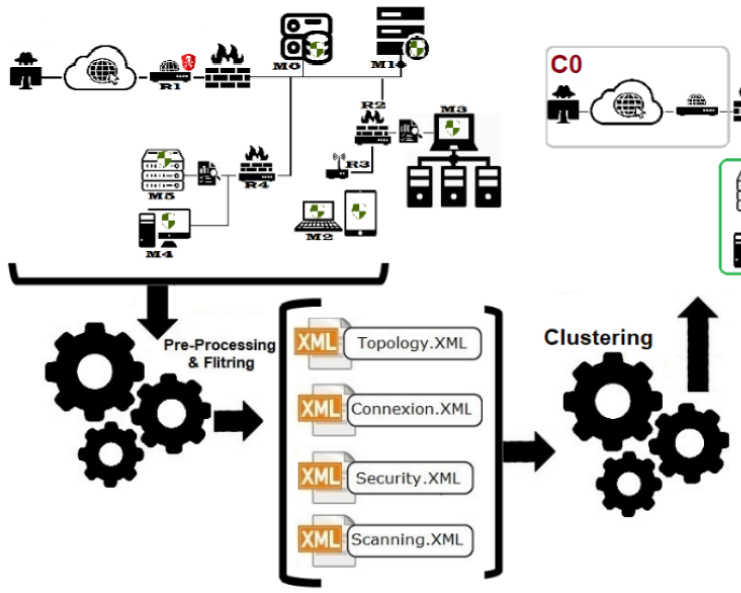
Belief Updates:

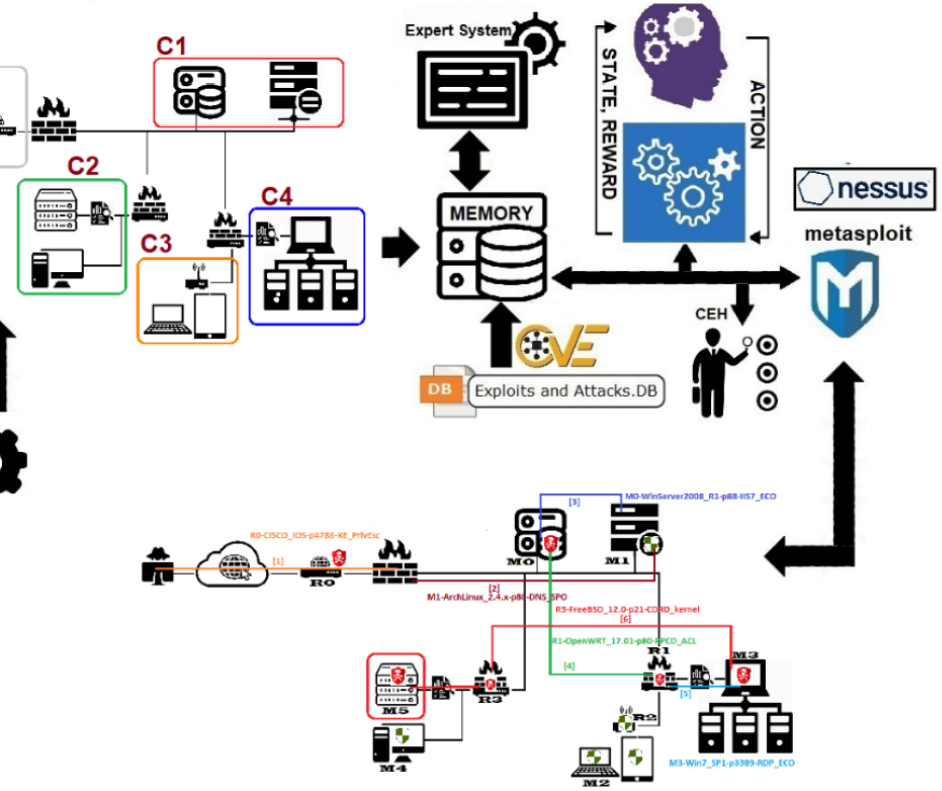
- Bayesian inference to adjust probabilities of being in S0, S1, or S2 based on traffic patterns.

RL in Offensive Security (Penetration Testing)

From: Reinforcement learning for efficient and effective penetration testing

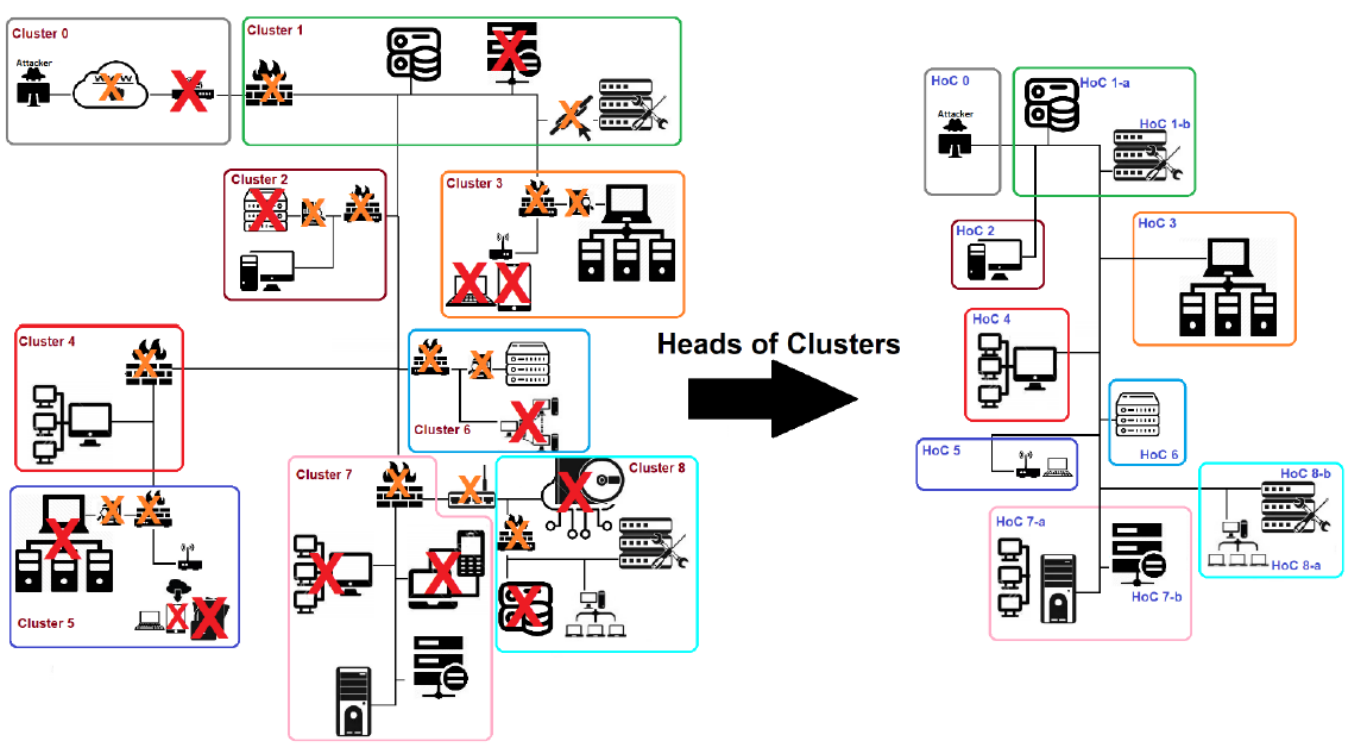


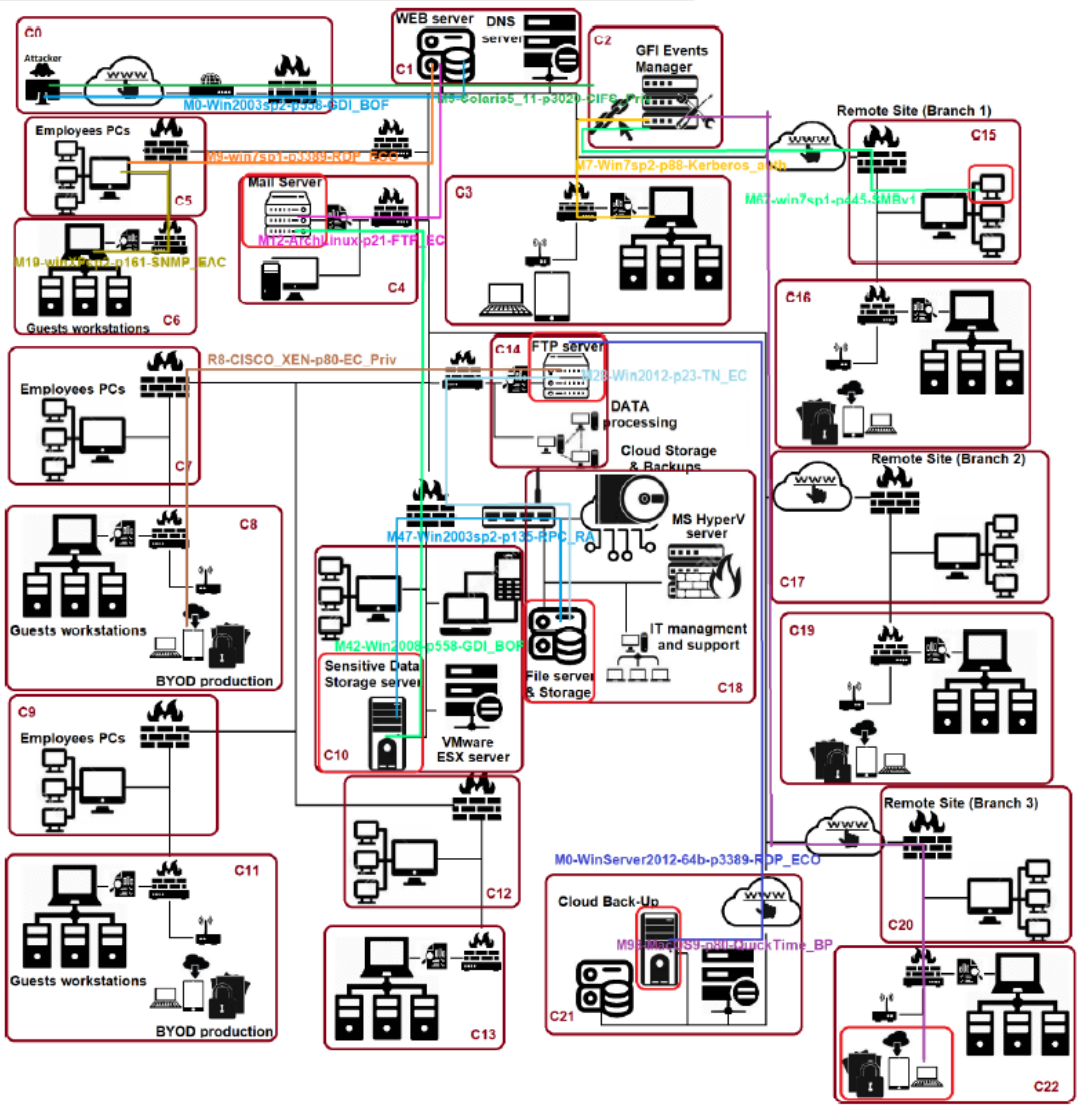




RL in Offensive Security (Penetration Testing) - Scalability

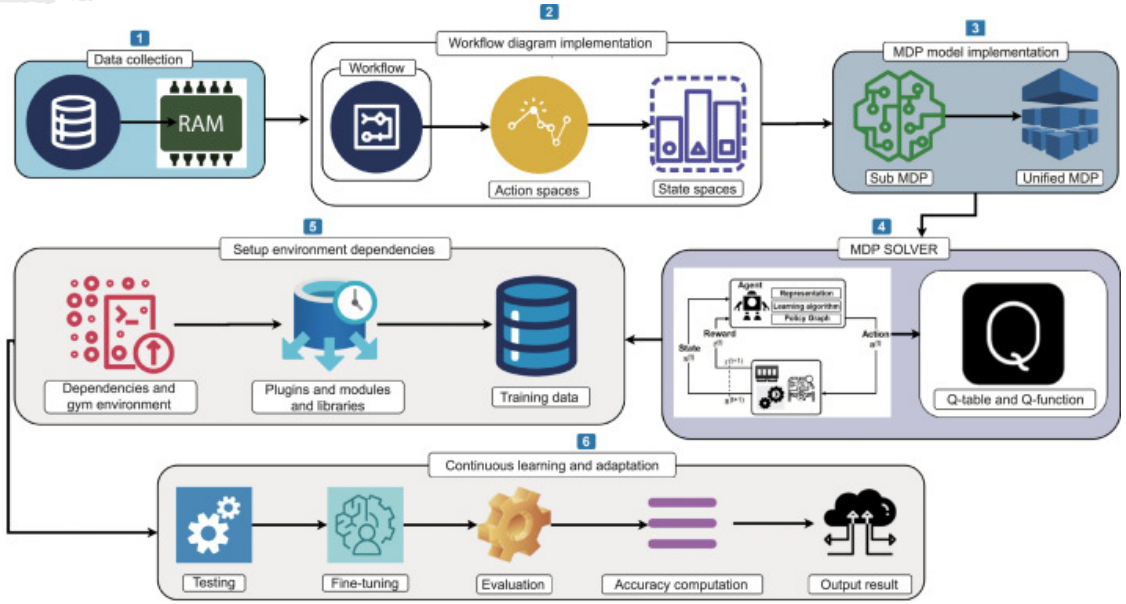
From: [Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks]

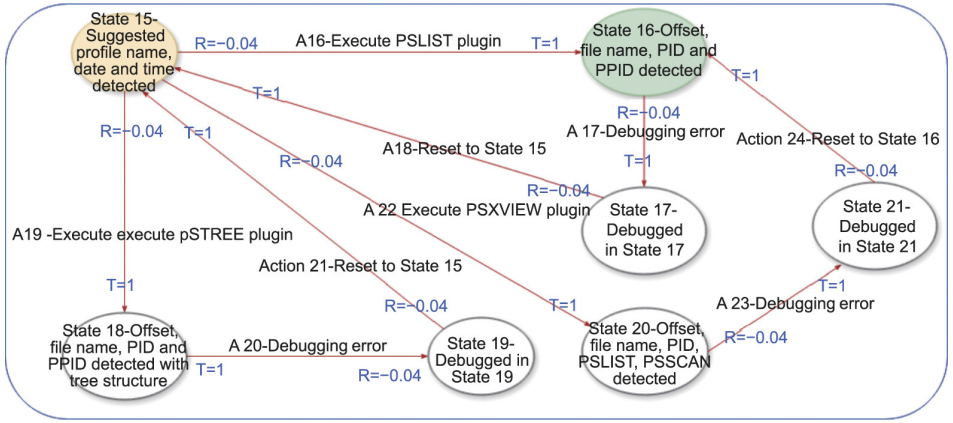




RL in Malware Cyber Incident Response

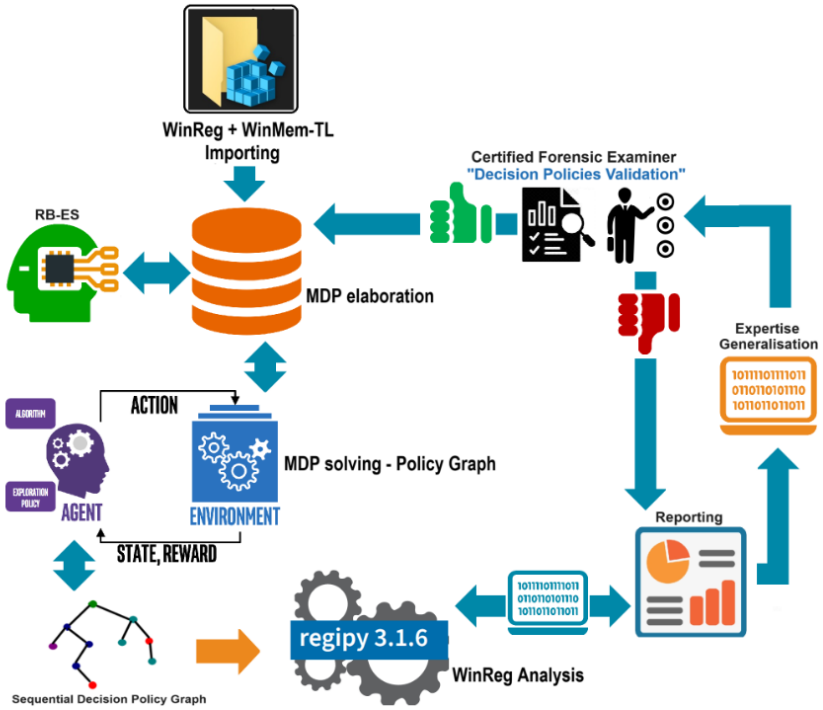
From: Reinforcement learning for an efficient and effective malware investigation during cyber Incident response

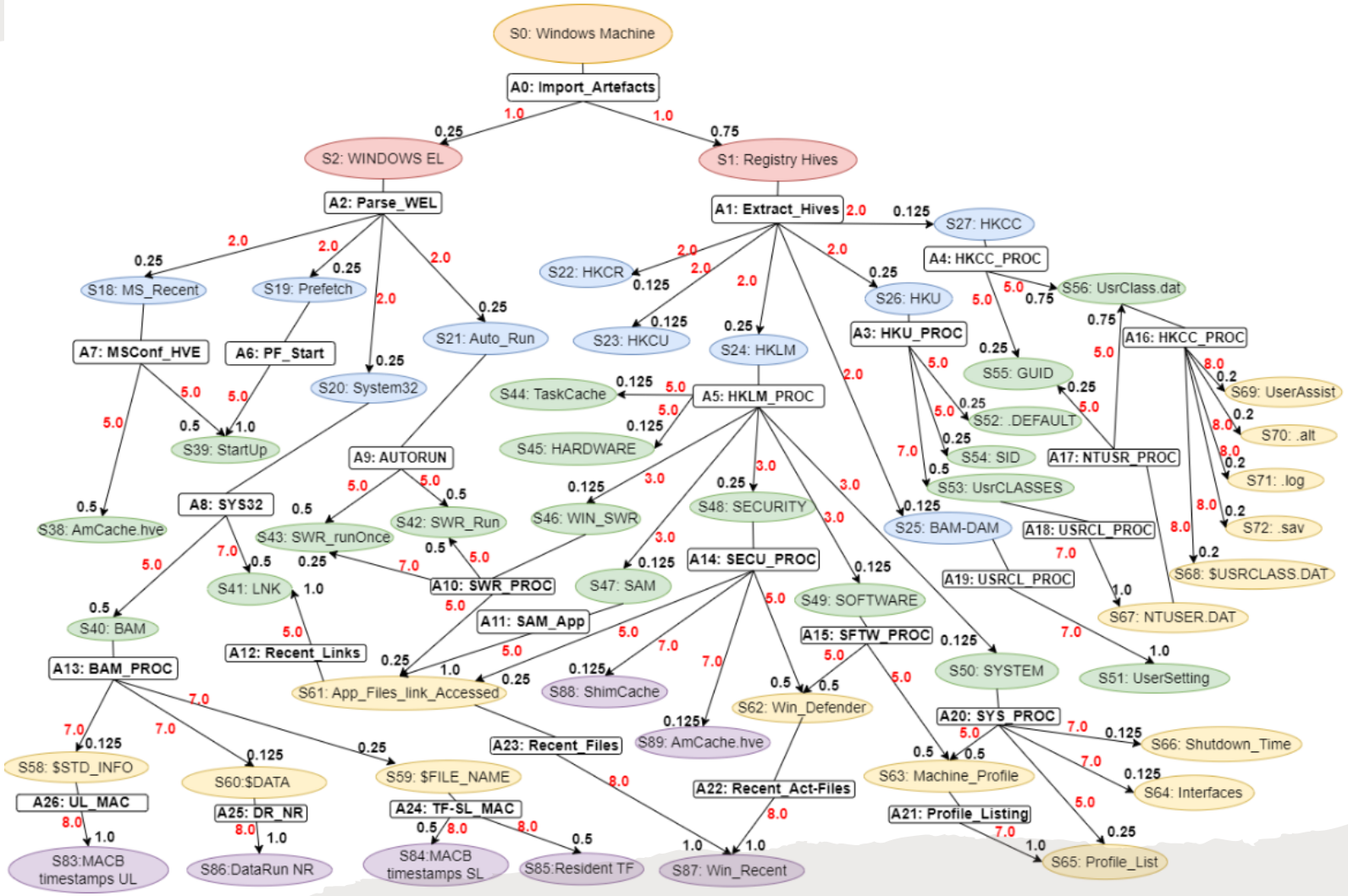




RL in Post Incident Investigation

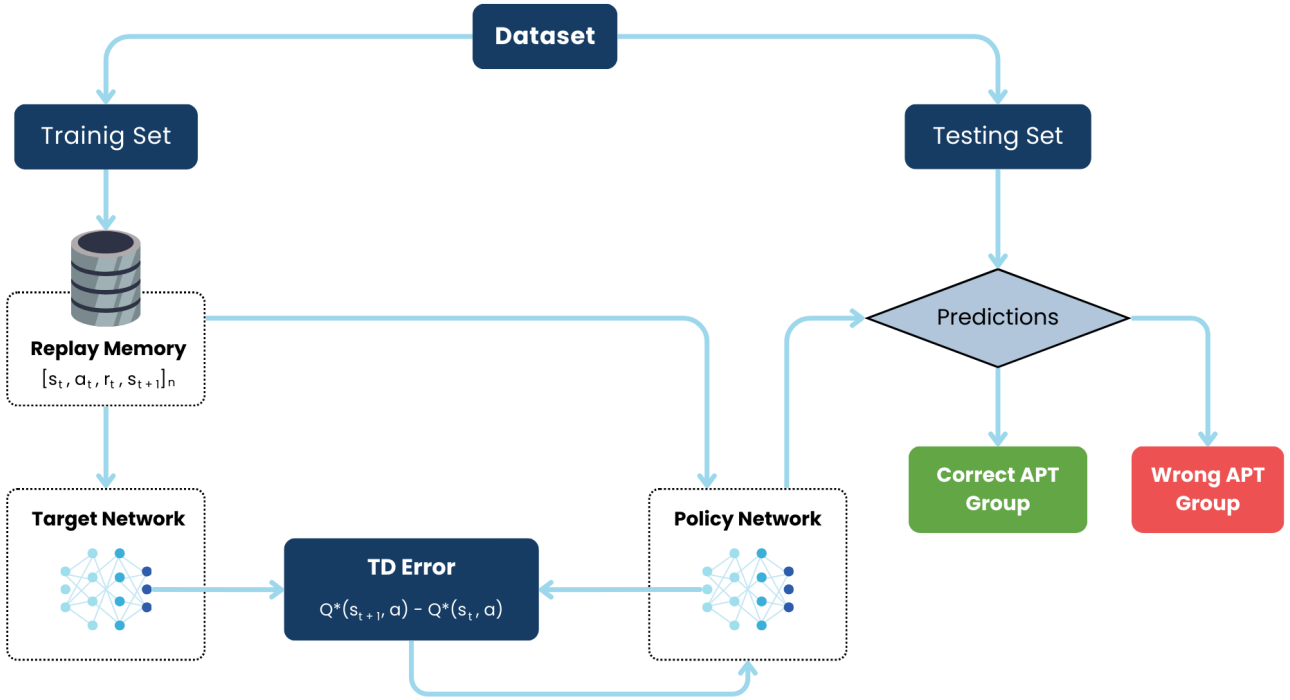
From: Leveraging Reinforcement Learning for an Efficient Automation of Windows Registry Analysis During Cyber Incident

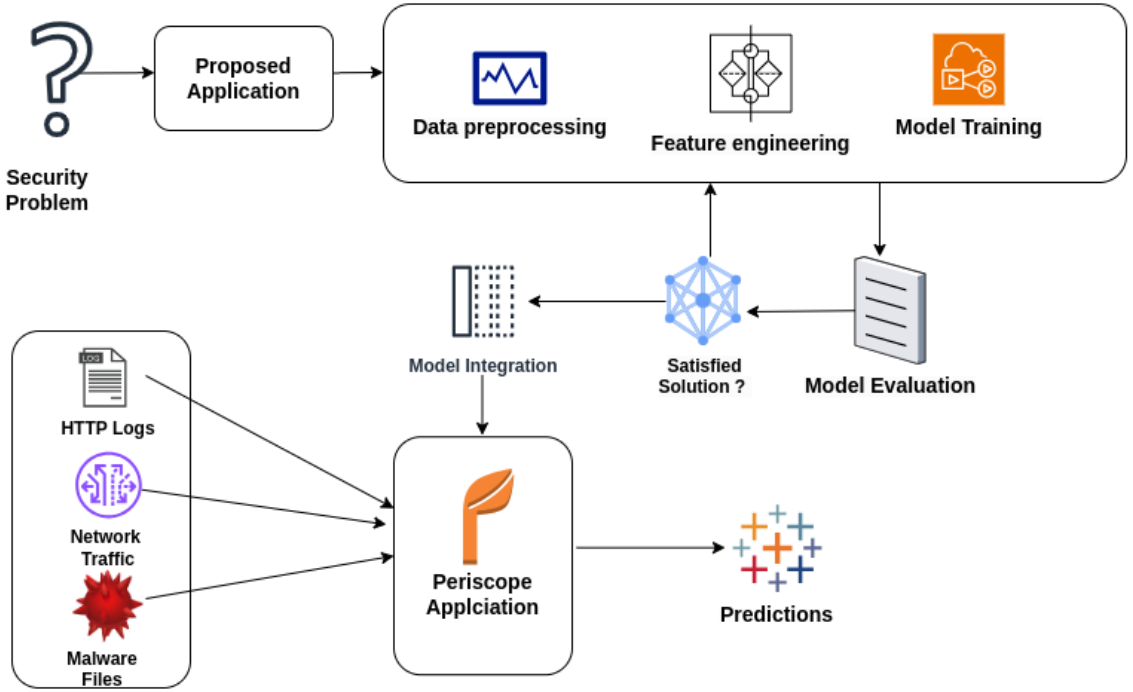




RL for ATP Attribution

From: Advanced Persistent Threats (APT) Attribution Using Deep Reinforcement Learning





Summary & Key Takeaways

- RL's potential in automating cybersecurity is clear
- RL tackle adversarial dynamics models (may require non-stationary MDPs).
- Importance of complete and realistic modelling (MDP/POMDP) in dynamic, uncertain environments
- Effective with MDP more complicated in MTD and PT where POMDP is needed
- Probabilistic belief states are required in POMDPs, to accurately capture the uncertainty from partial observability, allow for Bayesian updating
- Enhanced PERSEUS Algorithm: Improved using importance sampling for faster convergence and better efficiency
- RL in Cyber Security is a promising future research directions
- Emphasis on collaboration between theory and practice - Real-World Validation with live industry scenarios

Thank you!

Any Question?