KTH ROYAL INSTITUTE
OF TECHNOLOGY

Boosting Cyber Resilience with Human-in-the-loop AI

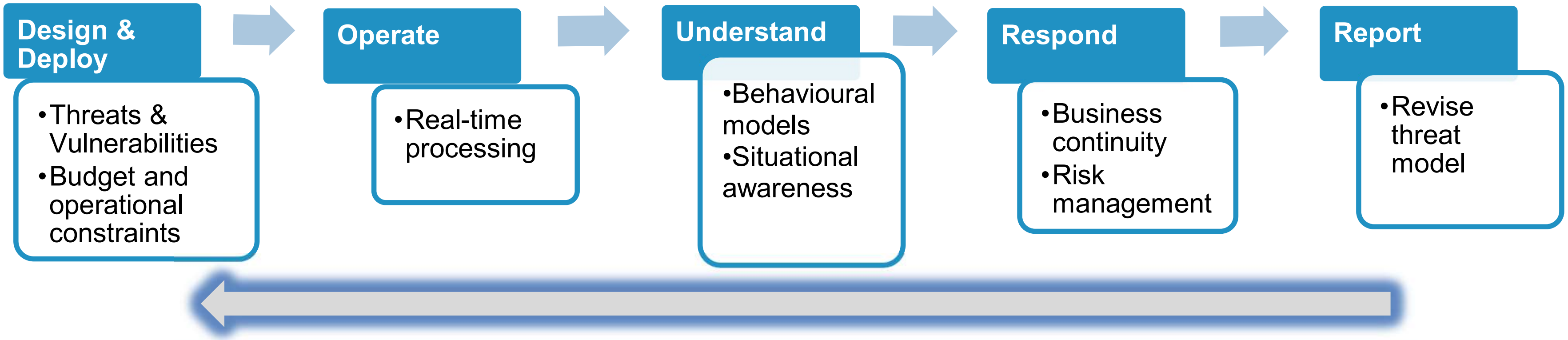
György Dán

IEEE CNS 2024 Workshop on Cyber Resilience

C3.ai Digital Transformation Institute

digital futures

The Needle in the Haystack?



Design & Deploy

- Threats & Vulnerabilities
- Budget and operational constraints

Operate

- Real-time processing

Understand

- Behavioural models
- Situational awareness

Respond

- Business continuity
- Risk management

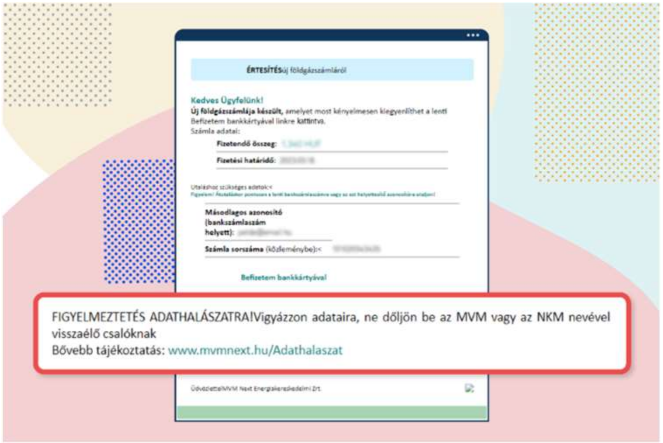
Report

- Revise threat model

AI Changes the Threat Model

AI-Powered Adversaries

- Social engineering
  - Target selection, deepfakes
- Phishing
  - Improved personalization, live communication at scale
- Vulnerability discovery
  - Hardware/software vulnerability analysis
- Autonomous malware



ÉRTESÍTÉSúj földgázszámláról

Kedves Ügyfelünk!
Új földgázszámlája készült, amelyet most kényelmesen kiegyenlíthet a lenti Befizetem bankkártyával linkre kattintva.
Számla adatai:

Fizetendő összeg: [redacted]
Fizetési határidő: [redacted]

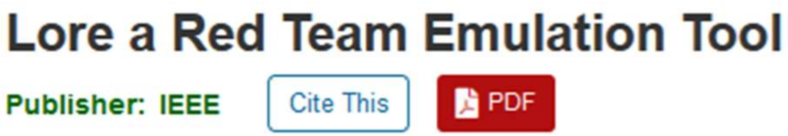
Utaláshoz szükséges adatok:<
Figyelem! Átutaláskor pontosan a lenti bankszámlaszámra vagy az azt helyettesítő azonosítóra utaljon!

Másodlagos azonosító (bankszámlaszám helyett): [redacted]
Számla sorszáma (közleménybe):< [redacted]

Befizetem bankkártyával

FIGYELMEZTETÉS ADATHALÁSZATRA!Vigyázzon adataira, ne dőljön be az MVM vagy az NKM nevével visszaélő csalóknak
Bővebb tájékoztatás: www.mvmnext.hu/Adathalaszat

Üdvözlettel/MVM Next Energiakereskedelmi Zrt.



Lore a Red Team Emulation Tool

Publisher: IEEE

New type of polymorphic fully autonomous malware uses AI

Technology News | August 2, 2023

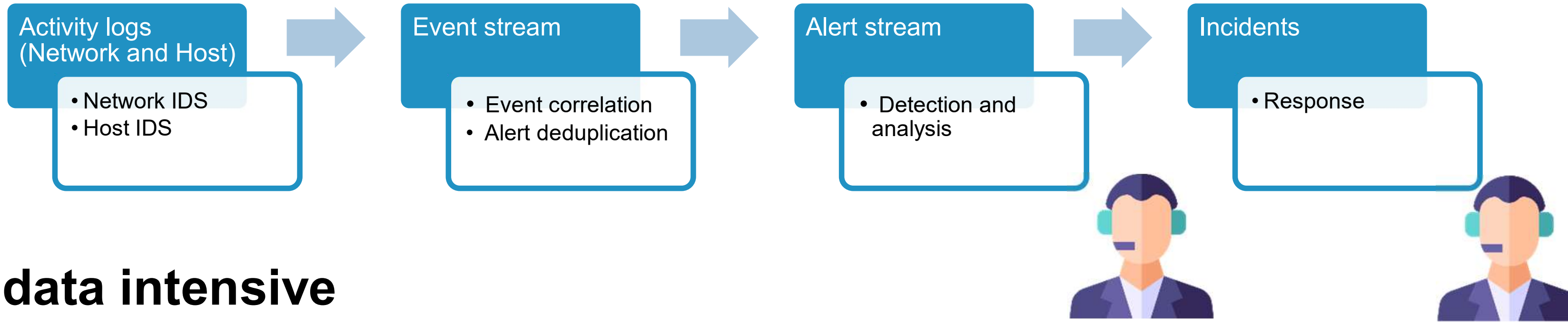
AI-Powered Cyber Resilience

- **From logs to incident response**

```
1,C625$@DOM1,U147@DOM1,C
625,C625,Negotiate,Batch

1,C653$@DOM1,SYSTEM@C65
3,C653,C653,Negotiate,Service

1,C660$@DOM1,SYSTEM@C66
0,C660,C660,Negotiate,Service
```



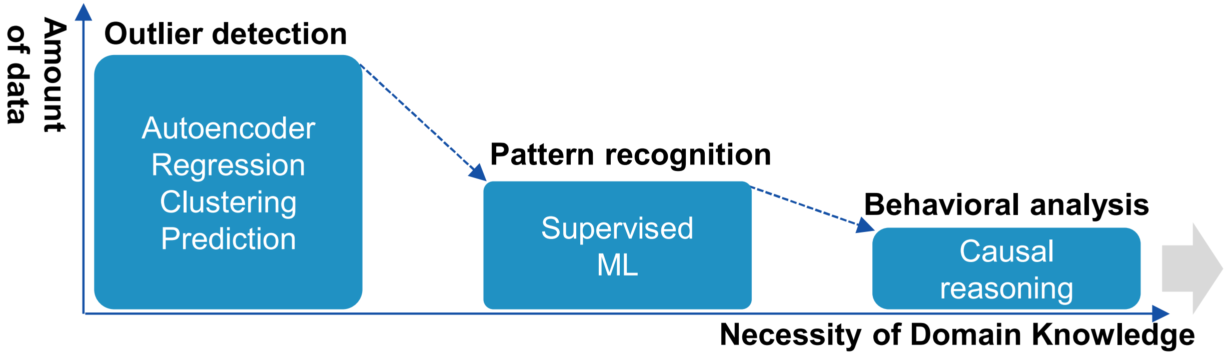
Activity logs (Network and Host)
- Network IDS
- Host IDS

Event stream
- Event correlation
- Alert deduplication

Alert stream
- Detection and analysis

Incidents
- Response

- **ML is data intensive**



Outlier detection

Autoencoder
Regression
Clustering
Prediction

Pattern recognition

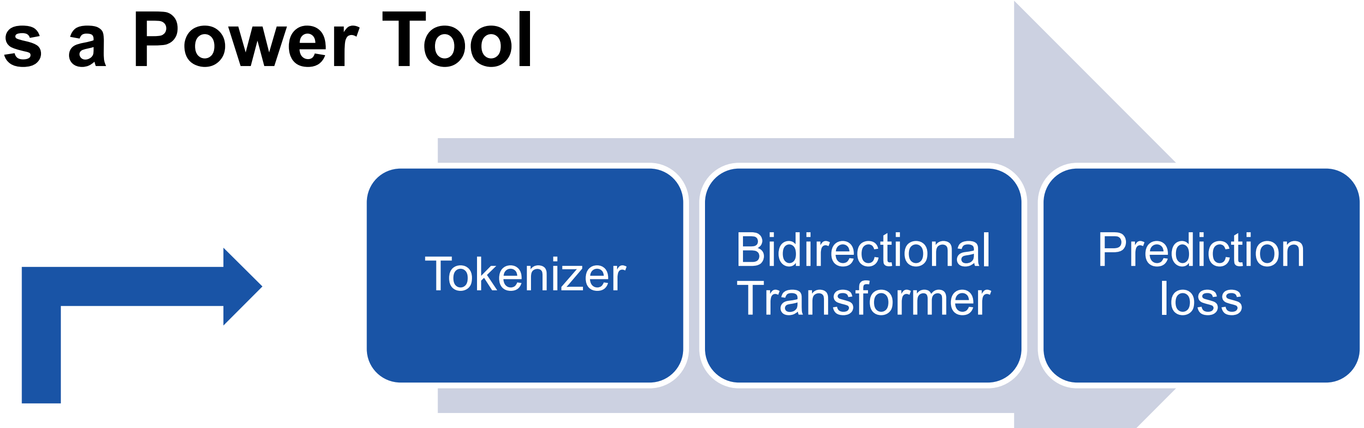
Supervised ML

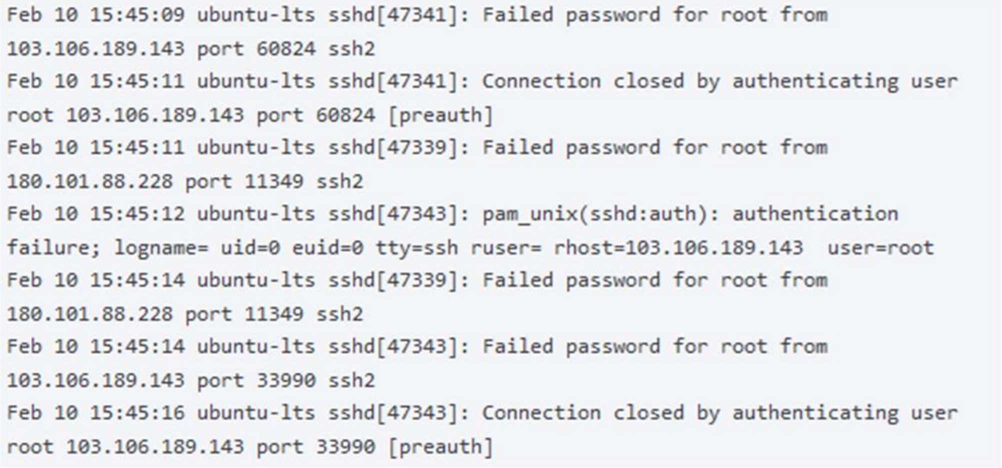
Behavioral analysis

Causal reasoning

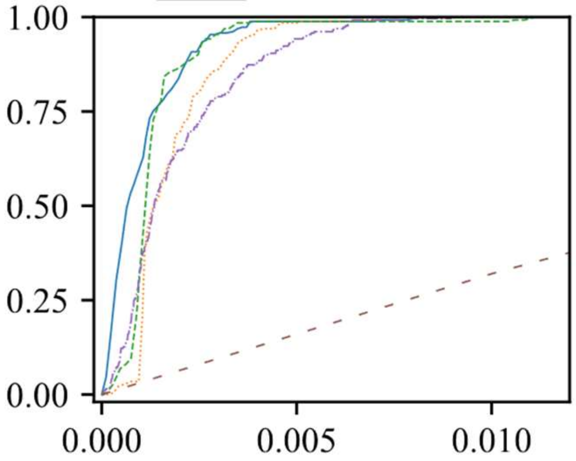
- **Business continuity**
- **Impact assessment**
  - Counterfactual reasoning
- **Interaction with humans**
  - Conversational AI

ML/AI as a Power Tool



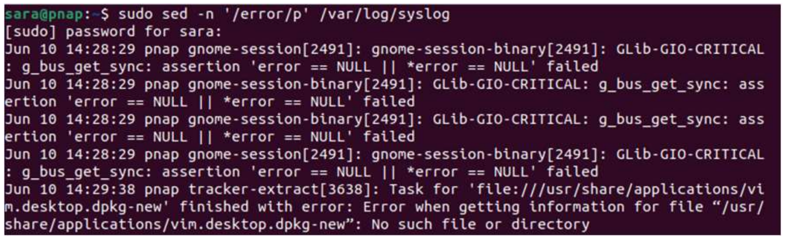


```
Feb 10 15:45:09 ubuntu-lts sshd[47341]: Failed password for root from
103.106.189.143 port 60824 ssh2
Feb 10 15:45:11 ubuntu-lts sshd[47341]: Connection closed by authenticating user
root 103.106.189.143 port 60824 [preauth]
Feb 10 15:45:11 ubuntu-lts sshd[47339]: Failed password for root from
180.101.88.228 port 11349 ssh2
Feb 10 15:45:12 ubuntu-lts sshd[47343]: pam_unix(sshd:auth): authentication
failure; logname= uid=0 euid=0 tty=ssh ruser= rhost=103.106.189.143  user=root
Feb 10 15:45:14 ubuntu-lts sshd[47339]: Failed password for root from
180.101.88.228 port 11349 ssh2
Feb 10 15:45:14 ubuntu-lts sshd[47343]: Failed password for root from
103.106.189.143 port 33990 ssh2
Feb 10 15:45:16 ubuntu-lts sshd[47343]: Connection closed by authenticating user
root 103.106.189.143 port 33990 [preauth]
```

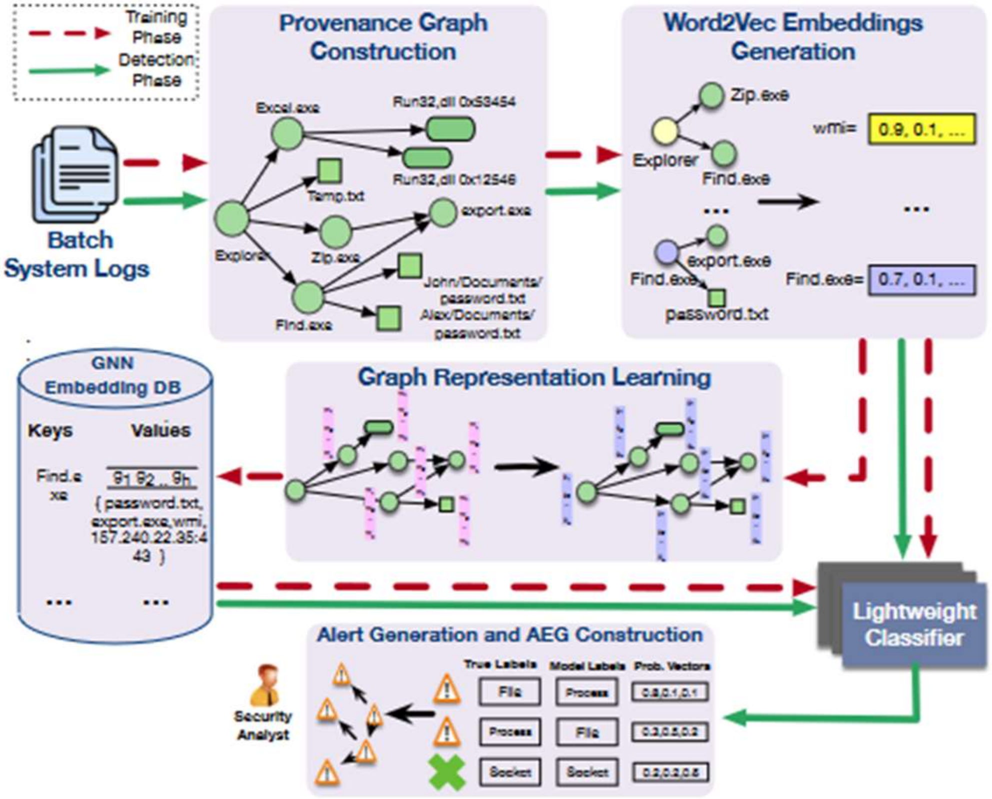


Gökstorp et al,``Anomaly Detection in Security Logs using Sequence Modeling,'' in Proc. of IFIP/IEEE NOMS, 2024

ML/AI as a Power Tool

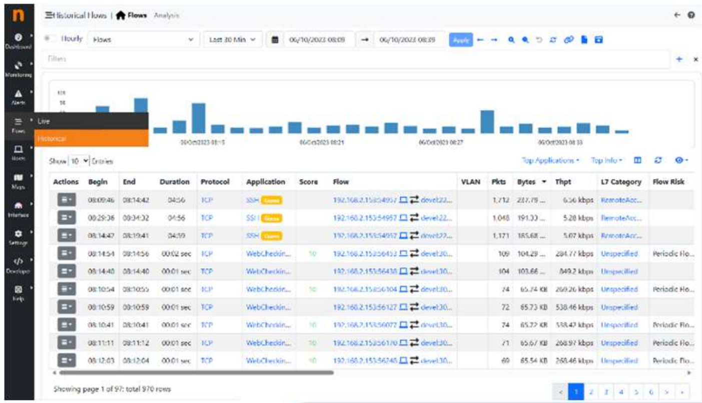


```
sara@pnap:~$ sudo sed -n '/error/p' /var/log/syslog
[sudo] password for sara:
Jun 10 14:28:29 pnap gnome-session[2491]: gnome-session-binary[2491]: GLib-GIO-CRITICAL
: g_bus_get_sync: assertion 'error == NULL || *error == NULL' failed
Jun 10 14:28:29 pnap gnome-session-binary[2491]: GLib-GIO-CRITICAL: g_bus_get_sync: ass
ertion 'error == NULL || *error == NULL' failed
Jun 10 14:28:29 pnap gnome-session-binary[2491]: GLib-GIO-CRITICAL: g_bus_get_sync: ass
ertion 'error == NULL || *error == NULL' failed
Jun 10 14:28:29 pnap gnome-session[2491]: gnome-session-binary[2491]: GLib-GIO-CRITICAL
: g_bus_get_sync: assertion 'error == NULL || *error == NULL' failed
Jun 10 14:29:38 pnap tracker-extract[3638]: Task for 'file:///usr/share/applications/vi
m.desktop.dpkg-new' finished with error: Error when getting information for file "/usr/
share/applications/vim.desktop.dpkg-new": No such file or directory
```



Rehman et al., "FLASH: A Comprehensive Approach to Intrusion Detection via Provenance Graph Representation Learning", in Proc. of IEEE S&P, 2024

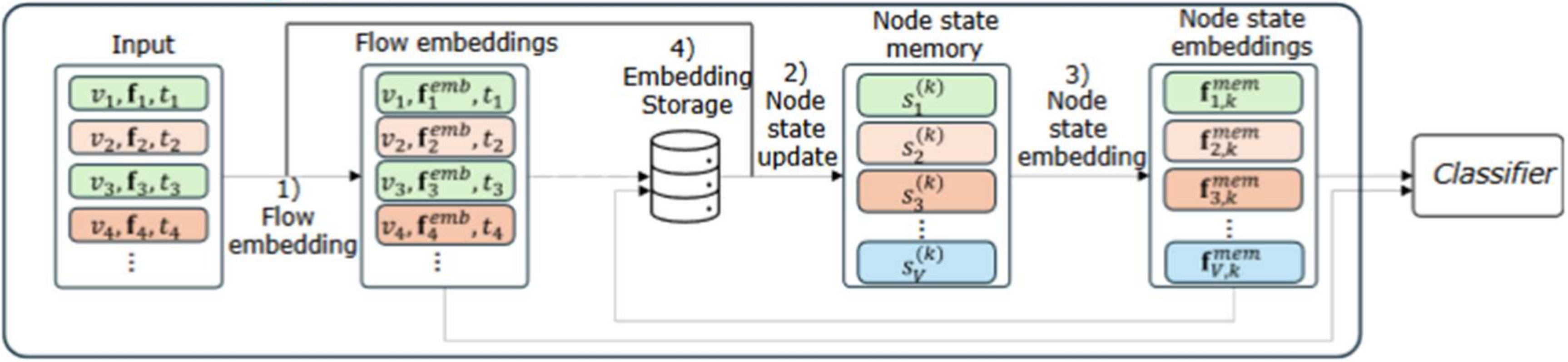
ML/AI as a Power Tool



Actions	Begin	End	Duration	Protocol	Application	Score	Flow	VLAN	Pkts	Bytes	Thpt	L7 Category	Flow Risk
☰	08:09:46	08:14:42	04:56	TCP	SSH		192.168.2.153:54957 ⇄ devel22...		1,712	237.79 ...	6.56 kbps	RemoteAcc...	
☰	08:29:36	08:34:32	04:56	TCP	SSH		192.168.2.153:54957 ⇄ devel22...		1,048	191.33 ...	5.28 kbps	RemoteAcc...	
☰	08:14:47	08:19:41	04:59	TCP	SSH		192.168.2.153:54957 ⇄ devel22...		1,171	185.68 ...	5.07 kbps	RemoteAcc...	
☰	08:14:54	08:14:56	00:02 sec	TCP	WebCheckin...	10	192.168.2.153:56453 ⇄ devel30...		109	104.29 ...	284.77 kbps	Unspecified	Periodic Flo...
☰	08:14:40	08:14:40	00:01 sec	TCP	WebCheckin...		192.168.2.153:56438 ⇄ devel30...		104	103.66 ...	849.2 kbps	Unspecified	
☰	08:10:54	08:10:55	00:01 sec	TCP	WebCheckin...	10	192.168.2.153:56104 ⇄ devel30...		74	65.74 KB	269.26 kbps	Unspecified	Periodic Flo...
☰	08:10:59	08:10:59	00:01 sec	TCP	WebCheckin...		192.168.2.153:56127 ⇄ devel30...		72	65.73 KB	538.46 kbps	Unspecified	
☰	08:10:41	08:10:41	00:01 sec	TCP	WebCheckin...	10	192.168.2.153:56077 ⇄ devel30...		74	65.72 KB	538.47 kbps	Unspecified	Periodic Flo...
☰	08:11:11	08:11:12	00:01 sec	TCP	WebCheckin...	10	192.168.2.153:56170 ⇄ devel30...		71	65.67 KB	268.97 kbps	Unspecified	Periodic Flo...
☰	08:12:03	08:12:04	00:01 sec	TCP	WebCheckin...	10	192.168.2.153:56248 ⇄ devel30...		69	65.54 KB	268.46 kbps	Unspecified	Periodic Flo...

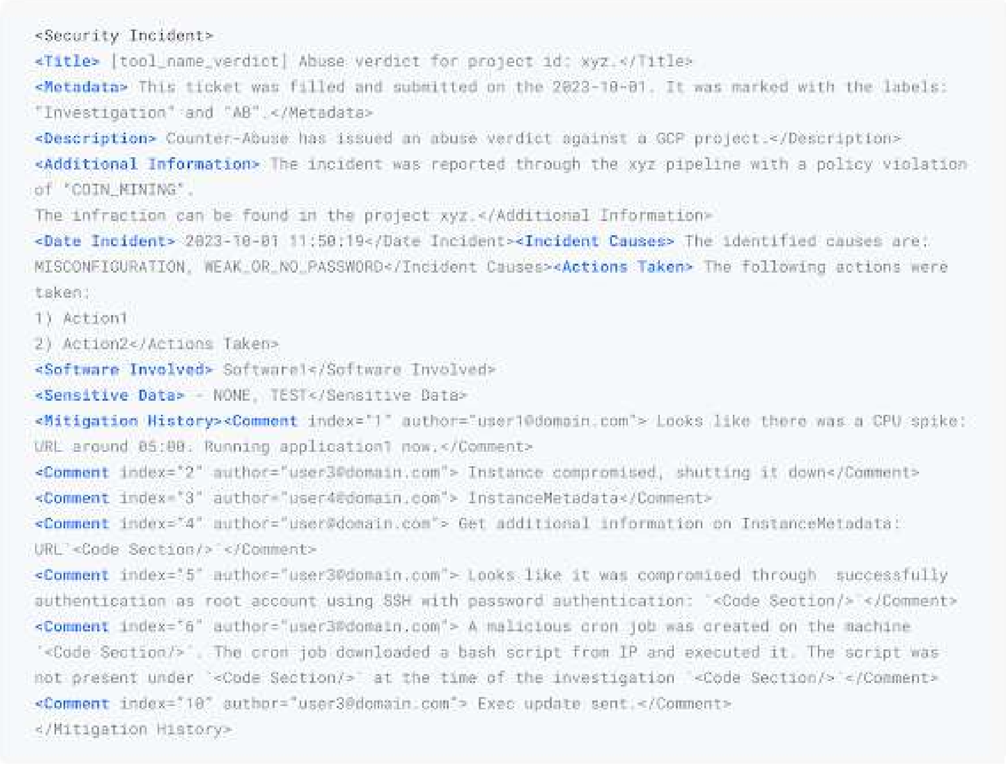
Showing page 1 of 97: total 970 rows

CCSTGN

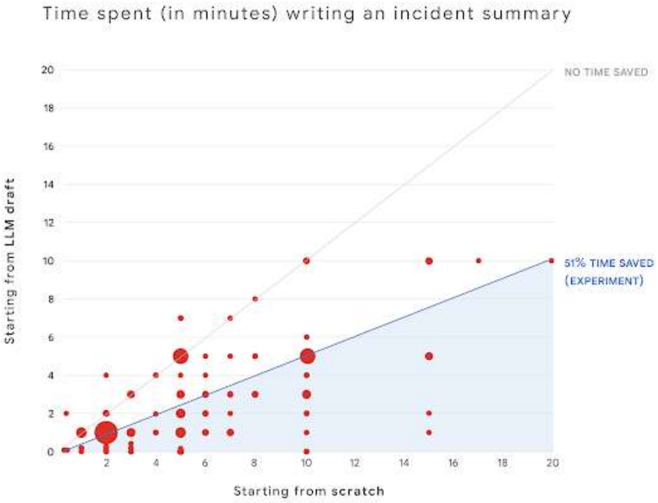


Santos et al., ``Channel-Centric Spatio-Temporal Graph Networks for Netw
based Intrusion Detection,'' in Proc. of IEEE CNS, 2024

ML/AI as a Power Tool

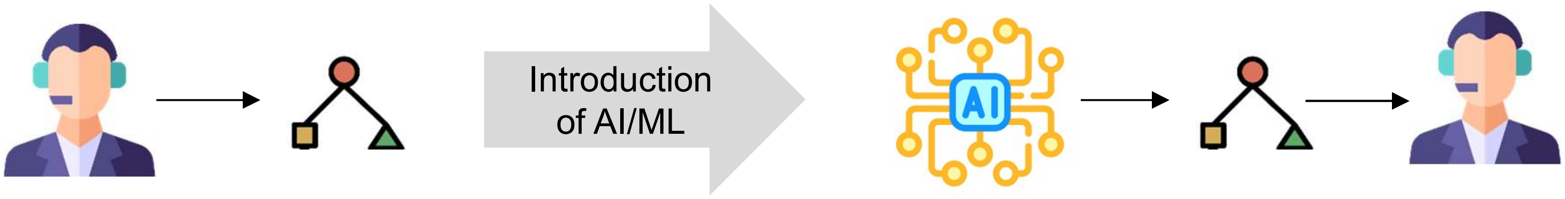


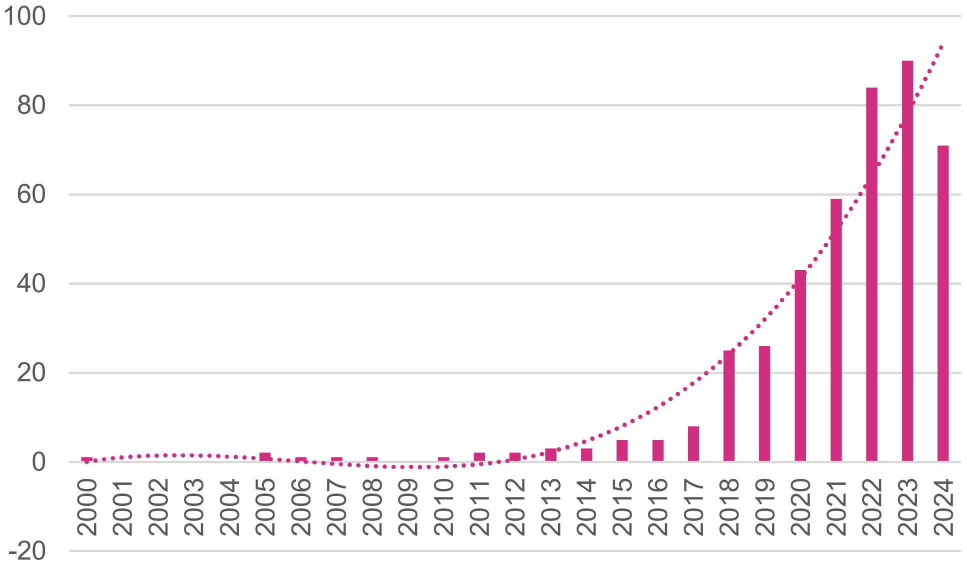
```
<Security Incident>
<Title> [tool_name_verdict] Abuse verdict for project id: xyz.</Title>
<Metadata> This ticket was filled and submitted on the 2023-10-01. It was marked with the labels:
"Investigation" and "AB".</Metadata>
<Description> Counter-Abuse has issued an abuse verdict against a GCP project.</Description>
<Additional Information> The incident was reported through the xyz pipeline with a policy violation
of "COIN_MINING".
The infraction can be found in the project xyz.</Additional Information>
<Date Incident> 2023-10-01 11:50:19</Date Incident><Incident Causes> The identified causes are:
MISCONFIGURATION, WEAK_OR_NO_PASSWORD</Incident Causes><Actions Taken> The following actions were
taken:
1) Action1
2) Action2</Actions Taken>
<Software Involved> Software1</Software Involved>
<Sensitive Data> - NONE, TEST</Sensitive Data>
<Mitigation History><Comment index="1" author="user1@domain.com"> Looks like there was a CPU spike:
URL around 05:00. Running application1 now.</Comment>
<Comment index="2" author="user3@domain.com"> Instance compromised, shutting it down</Comment>
<Comment index="3" author="user4@domain.com"> InstanceMetadata</Comment>
<Comment index="4" author="user@domain.com"> Get additional information on InstanceMetadata:
URL`<Code Section/>`</Comment>
<Comment index="5" author="user3@domain.com"> Looks like it was compromised through  successfully
authentication as root account using SSH with password authentication: `<Code Section/>`</Comment>
<Comment index="6" author="user3@domain.com"> A malicious cron job was created on the machine
`<Code Section/>`. The cron job downloaded a bash script from IP and executed it. The script was
not present under `<Code Section/>` at the time of the investigation `<Code Section/>`</Comment>
<Comment index="10" author="user3@domain.com"> Exec update sent.</Comment>
</Mitigation History>
```



– https://security.googleblog.com/2024/04/accelerating-incident-response-using.html

Autonomous Cyber Defense


Introduction of AI/ML

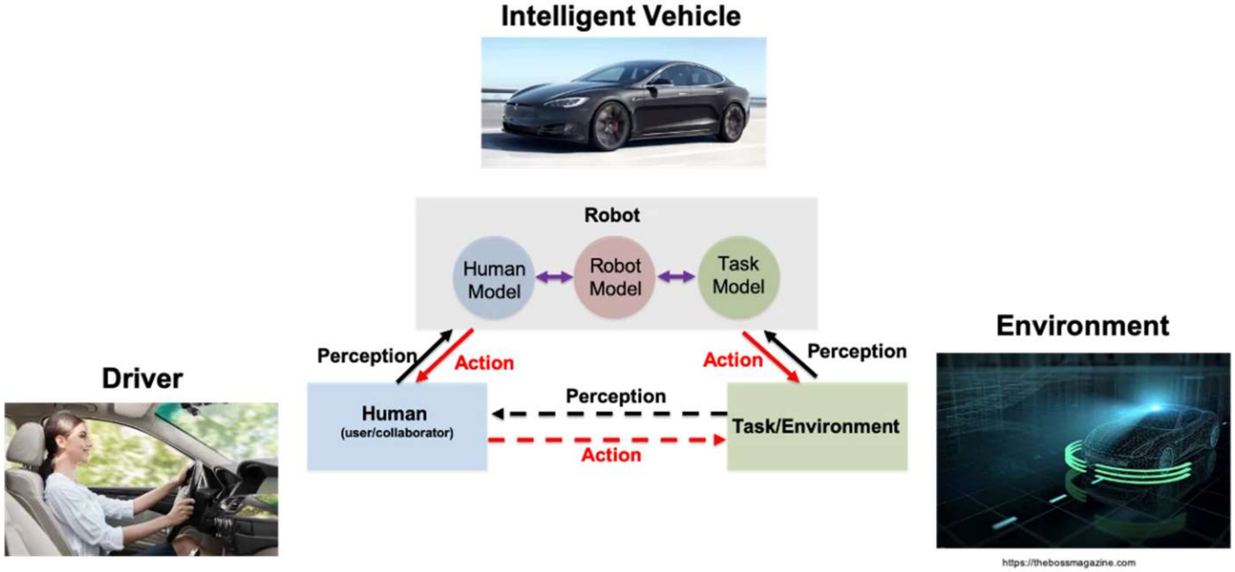


Source: https://github.com/Limmen/awesome-rl-for-cybersecurity

- **Operational requirements**
  - Auditable
  - Controllable
  - Transferable/Adaptive
  - Secure
  - Observable/Explainble

Human-in-the-loop AI for Security

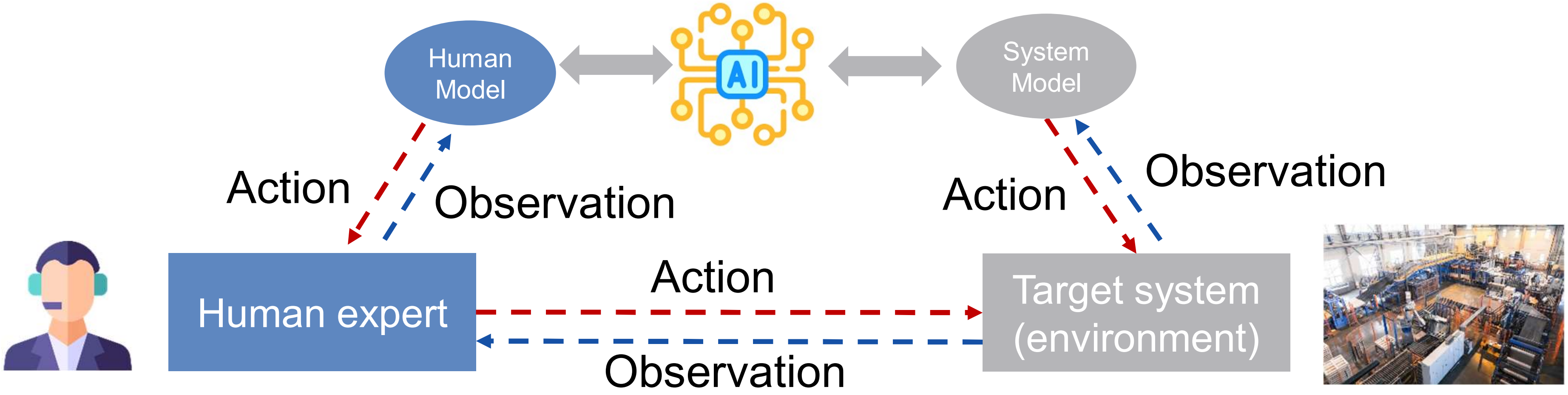
- **AI/ML complementing human decision making**
  - Reduced response time
  - Higher accuracy



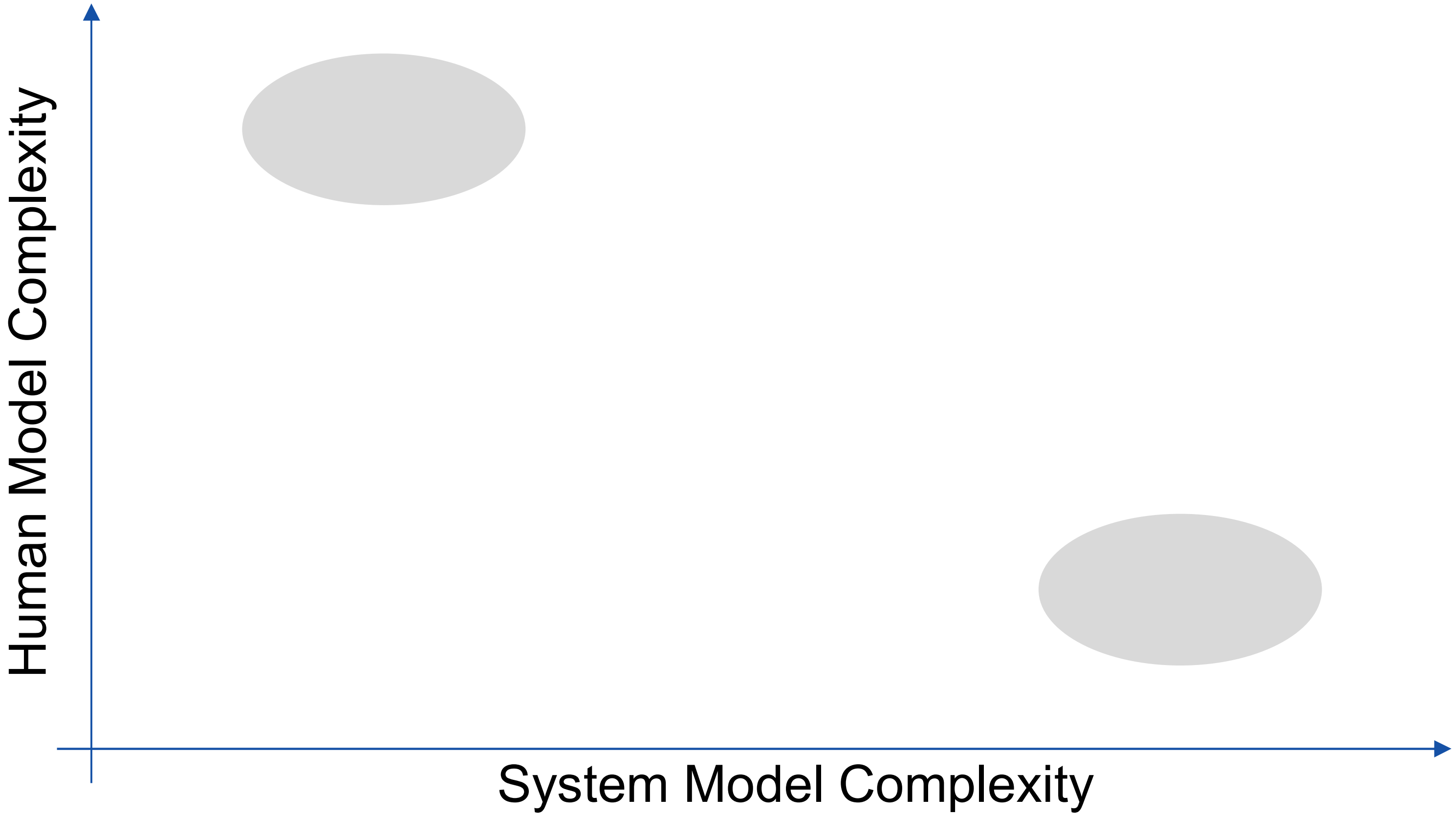
https://arm.stanford.edu/research/leveraging-human-intent-shared-autonomy

Human-in-the-loop AI for Security

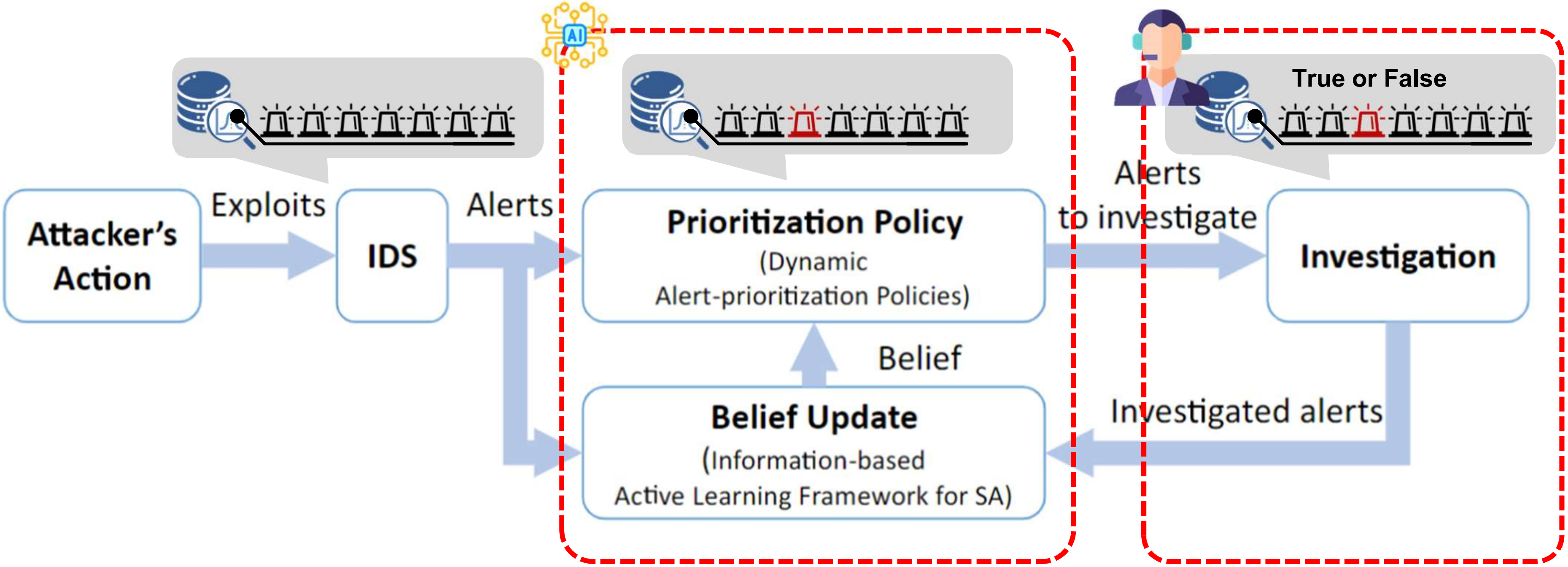
- **AI/ML complementing human decision making**
  - Reduced response time
  - Higher accuracy



Framework Design Space

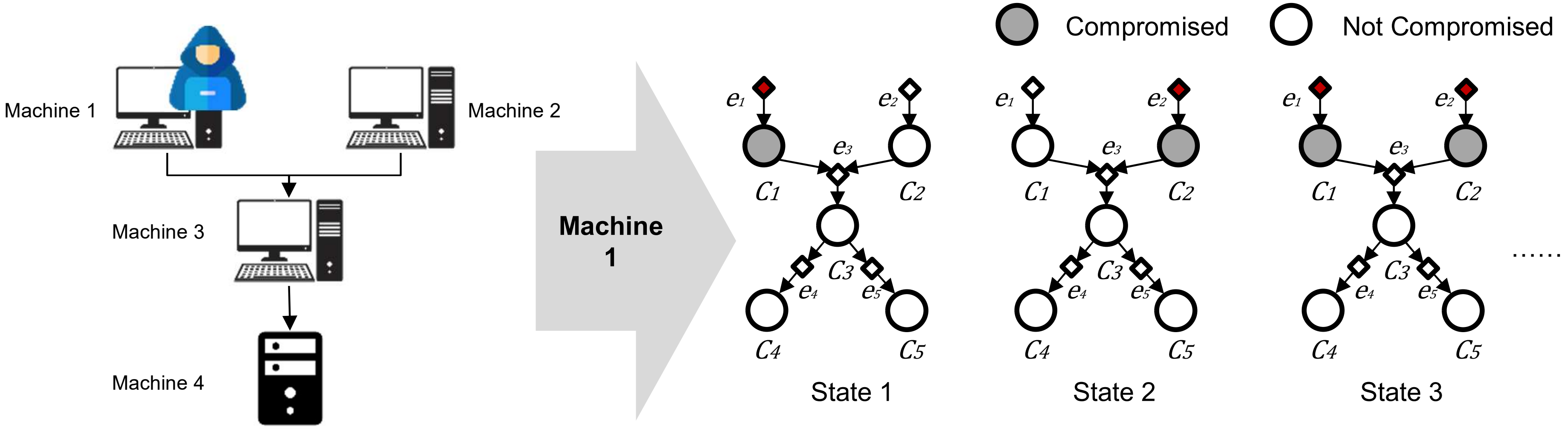


Human-in-the-Loop AI Framework



State in Cyber Security

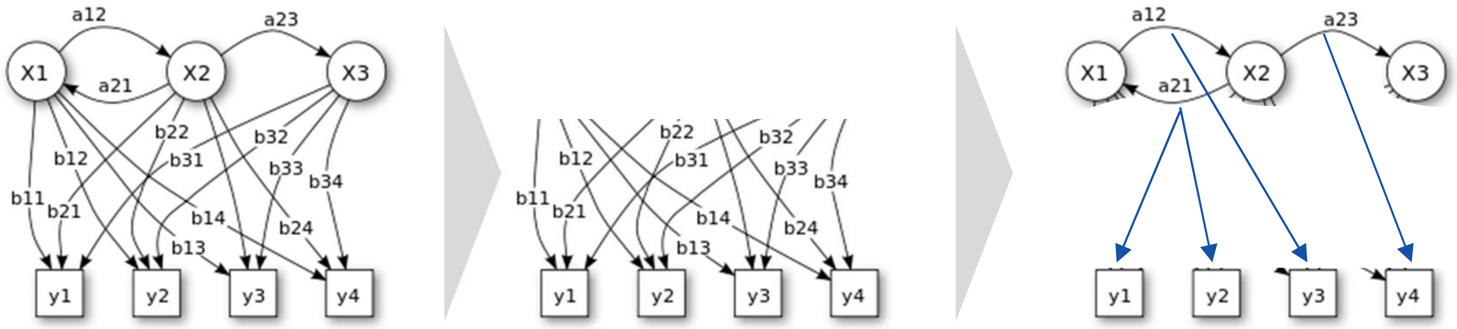
- Attack tree: Hypergraph of conditions and exploits
- Attack state: the set of conditions/privileges the attacker gained



- States and transitions → Markov model

Problem of Partial Observability

- Security state is not visible to the defender
  - Attacker activity can trigger alerts



X — states
y — possible observations
a — state transition probabilities
b — output probabilities

- Hidden Markov model

System model – Security state

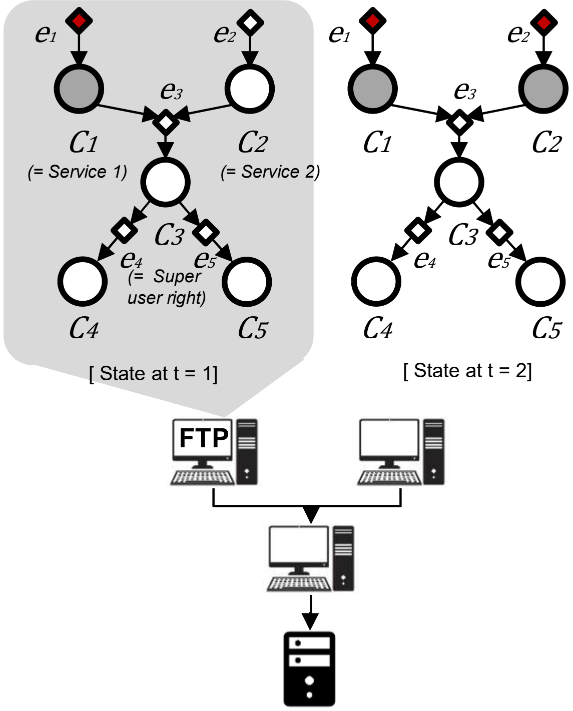
- **Time is slotted**
- **Attack Hypergraph**
  - Nodes: conditions (access privilege, etc)
  - Hyperedges: exploits

$$\mathcal{H} = (\mathcal{N}, \mathcal{E}) \text{ where } \mathcal{N} = \{c_1, ..., c_{n_c}\}, \text{ and } \mathcal{E} = \{e_1, ..., e_{n_e}\}$$

- **Security state: set of enabled conditions**

$$s_1 = \{c_1\} \qquad s_2 = \{c_1, c_2\}$$

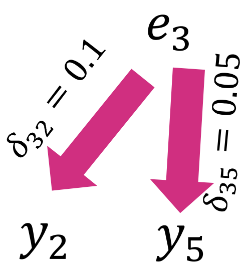
- **Example**
  - c1: wu-ftpd 2.5 running on host
  - c2: ftp server remotely accessible
  - e3: CVE-1999-0878
  - c3: Root privilege on host



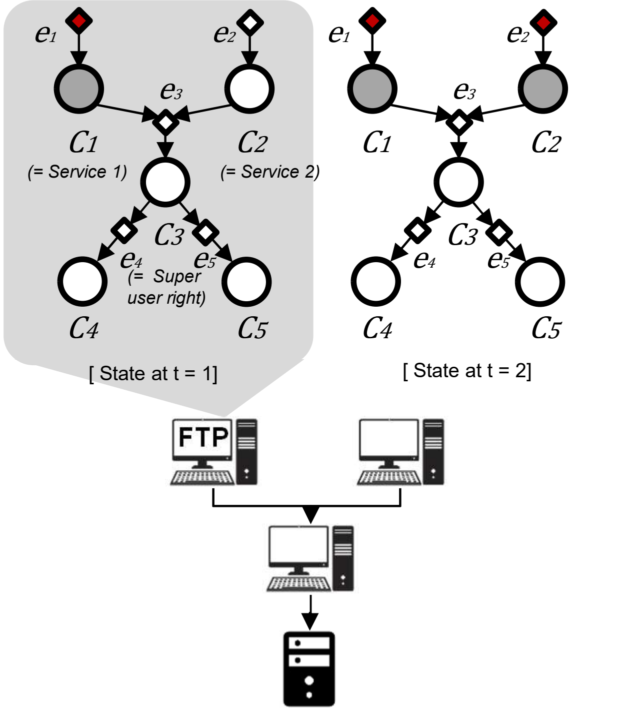
Kim et al, "An Active Learning Approach to Dynamic Alert Prioritization for Real-time Situational Awareness" Proc. of IEEE CNS, 2022

Attacker model

- **Attacker chooses exploits independently**
  - Probability of choosing exploit e_i: α_{e_i}
  - Probability that exploit e_i succeeds: β_{e_i}
- **If exploit e_i is used**
  - Generates alert a with probability δ_{ia}



- **False positive** with probability ζ_a
- **Alert vector** $Y_t = (y_1, \dots, y_{n_z})$



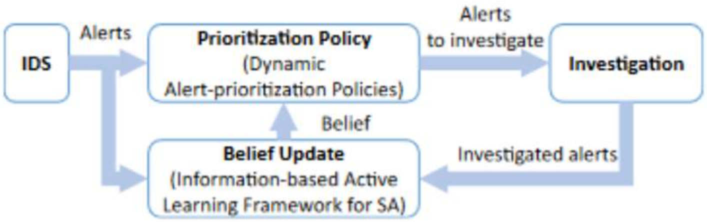
Kim et al, "An Active Learning Approach to Dynamic Alert Prioritization for Real-time Situational Awareness" Proc. of IEEE CNS, 2022

Defender model

- **Observation at time *t*:** Y_t (alert vector)
- **Action:**
  - Inspect up to *l* alerts in Y_t
  - Inspecting alert y_t^a results in modified alert $\hat{y}_t^a$
- **Human model:** Investigation error probability ω
- **Belief about security state**

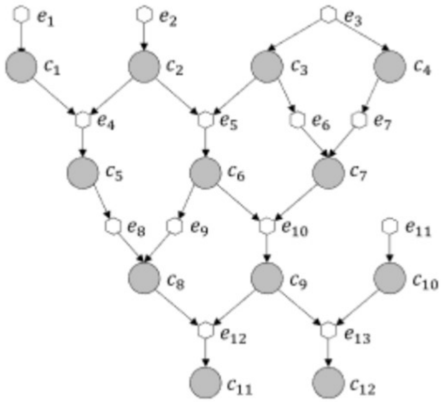
$$\pi_t = \begin{bmatrix} \pi_t^{1,1} & \pi_t^{1,2} & \cdots & \pi_t^{1,n_a} \\ \pi_t^{2,1} & \pi_t^{2,2} & \cdots & \pi_t^{2,n_a} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_t^{n_s,1} & \pi_t^{n_s,2} & \cdots & \pi_t^{n_s,n_a} \end{bmatrix} \in \Delta(\mathcal{S} \times \Phi)$$

- **Cost:** State estimation error $J^\kappa = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \gamma^t MSE(\pi_t^\kappa, s_t^\kappa)$
- **Optimal policy:** $\kappa^* \in \arg\min_{\kappa \in \mathcal{K}} J^\kappa$



	Investigation outcome	
Ground truth	TP	FP
TP	$1-\omega$	ω
FP	ω	$1-\omega$

Active learning for alert prioritization



- **In practice the state is unknown → cannot calculate MSE**

> *Use **belief uncertainty as a proxy** for the MSE.*
> *Intuition: Low uncertainty is likely to imply an accurate belief*

- **Proposed candidate policies**
  - Max-entropy

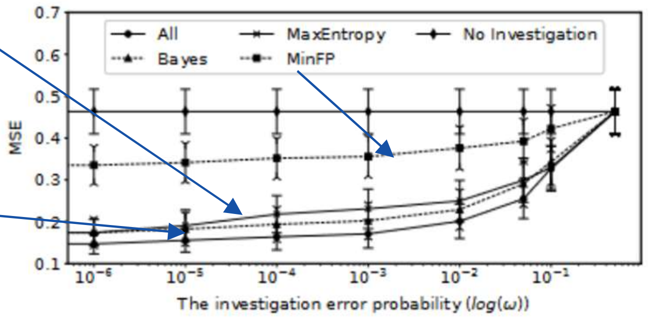
    → Investigate the alert v that decreases the entropy most

    $$\min_{v} \mathrm{H}(S_{t+1} = s_{i'}, \Phi_{t+1} = \phi_{l'} | V_{t+1} = v, Y_{t+1} = y_n, \Pi_t = \pi_t)$$

  - Bayes factor policy

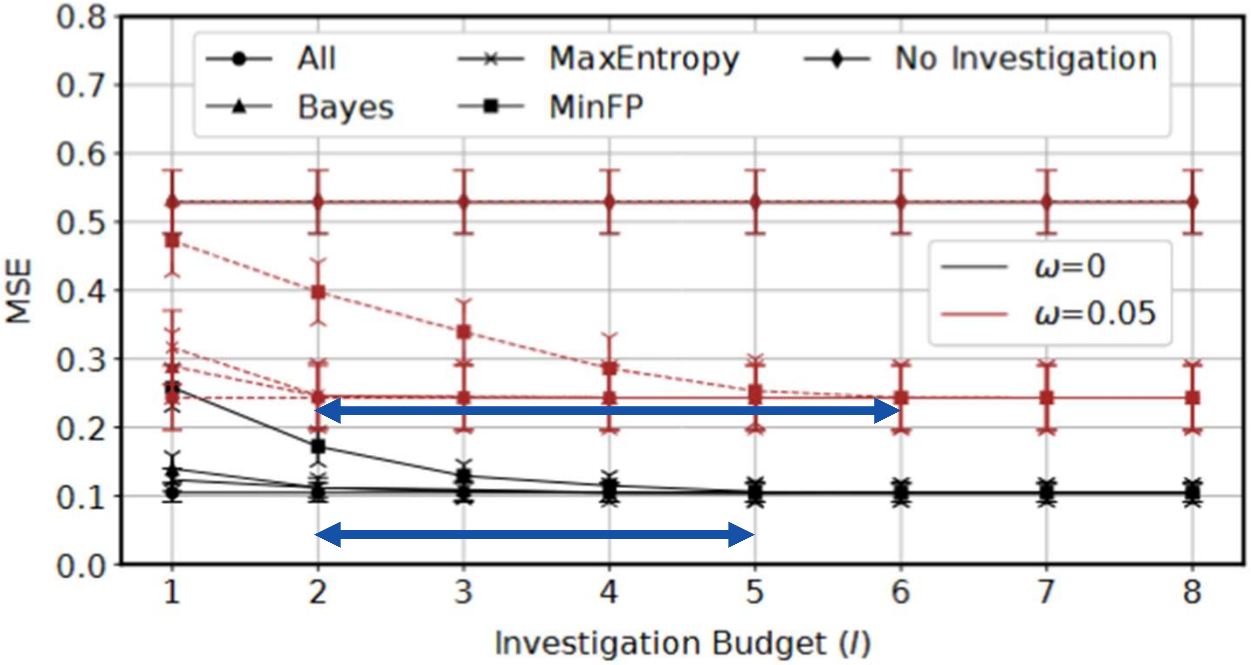
    → Investigate the most ambiguous alert
    (alert probability without false positives vs. false positive rate)

    $$K^a = \frac{P(Y^a_{t+1} = 1 \mid Y^{-a}_{t+1} = y_n^{-a}, \Pi_t = \pi_t)|_{\zeta_a = 0}}{\zeta_a},$$



Kim et al, "An Active Learning Approach to Dynamic Alert Prioritization for Real-time Situational Awareness" Proc. of *IEEE CNS,* 2022

System Level Benefit

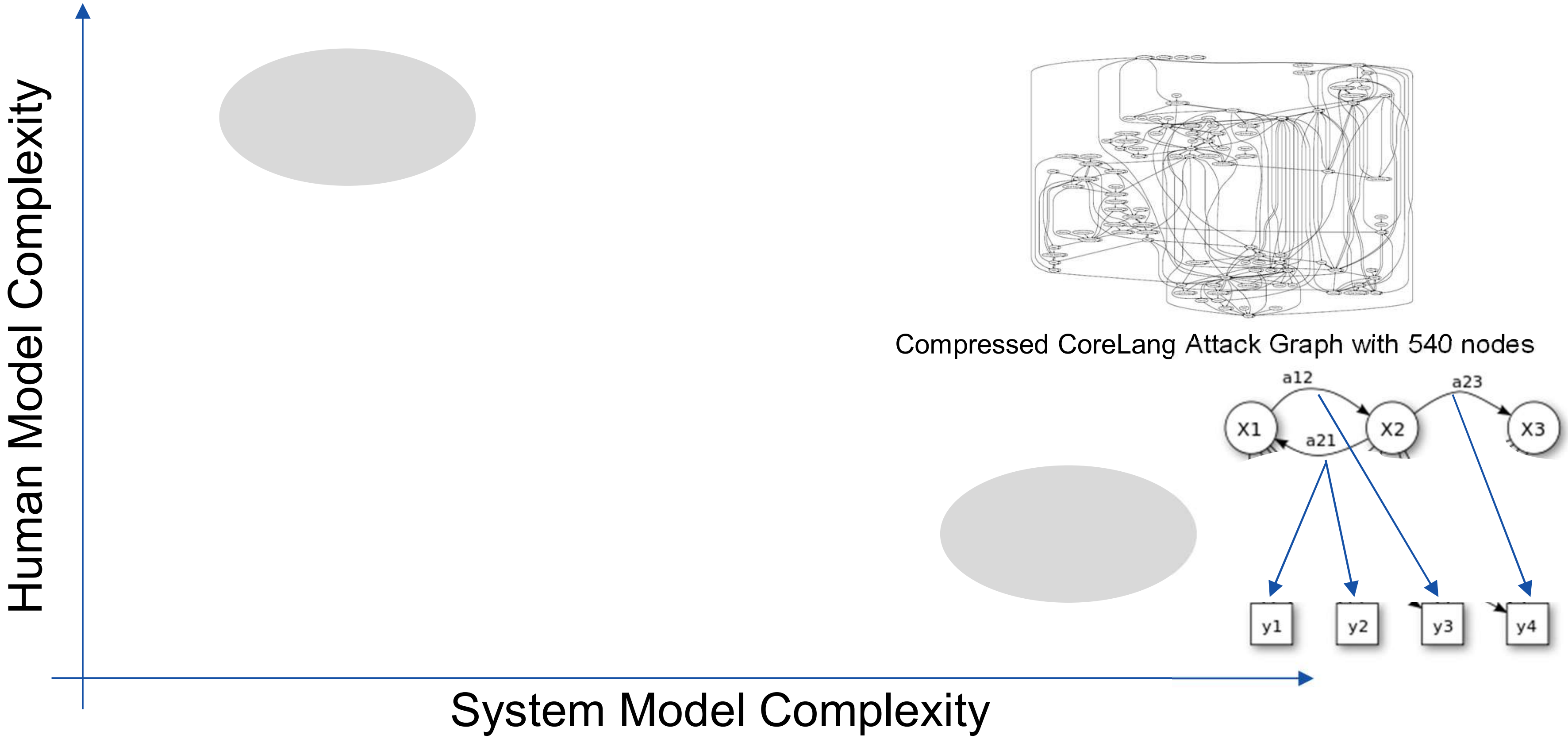


Less human effort needed

Less skilled experts needed

Kim et al, "An Active Learning Approach to Dynamic Alert Prioritization for Real-time Situational Awareness" Proc. of *IEEE CNS,* 2022

Framework Design Space



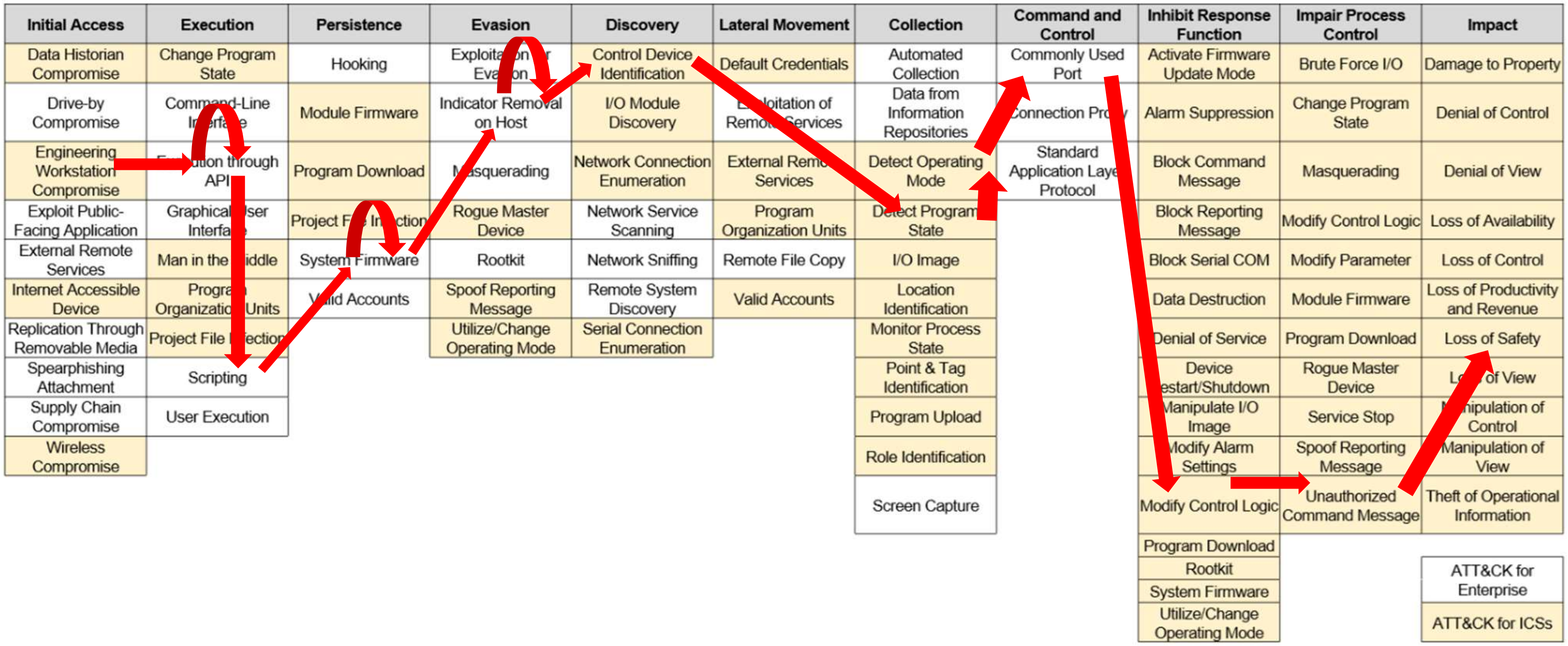
Human Model Complexity

System Model Complexity

Compressed CoreLang Attack Graph with 540 nodes

Katsikeas et al. “An attack simulation language for the IT domain,” in *Proc. of Int. Workshop on Graphical Models for Security*, pp. 67–86, 2020.

MITRE ATT&CK Model



Initial Access	Execution	Persistence	Evasion	Discovery	Lateral Movement	Collection	Command and Control	Inhibit Response Function	Impair Process Control	Impact
Data Historian Compromise	Change Program State	Hooking	Exploitation for Evasion	Control Device Identification	Default Credentials	Automated Collection	Commonly Used Port	Activate Firmware Update Mode	Brute Force I/O	Damage to Property
Drive-by Compromise	Command-Line Interface	Module Firmware	Indicator Removal on Host	I/O Module Discovery	Exploitation of Remote Services	Data from Information Repositories	Connection Proxy	Alarm Suppression	Change Program State	Denial of Control
Engineering Workstation Compromise	Execution through API	Program Download	Masquerading	Network Connection Enumeration	External Remote Services	Detect Operating Mode	Standard Application Layer Protocol	Block Command Message	Masquerading	Denial of View
Exploit Public-Facing Application	Graphical User Interface	Project File Infection	Rogue Master Device	Network Service Scanning	Program Organization Units	Detect Program State		Block Reporting Message	Modify Control Logic	Loss of Availability
External Remote Services	Man in the Middle	System Firmware	Rootkit	Network Sniffing	Remote File Copy	I/O Image		Block Serial COM	Modify Parameter	Loss of Control
Internet Accessible Device	Program Organization Units	Valid Accounts	Spoof Reporting Message	Remote System Discovery	Valid Accounts	Location Identification		Data Destruction	Module Firmware	Loss of Productivity and Revenue
Replication Through Removable Media	Project File Infection		Utilize/Change Operating Mode	Serial Connection Enumeration		Monitor Process State		Denial of Service	Program Download	Loss of Safety
Spearphishing Attachment	Scripting					Point & Tag Identification		Device Restart/Shutdown	Rogue Master Device	Loss of View
Supply Chain Compromise	User Execution					Program Upload		Manipulate I/O Image	Service Stop	Manipulation of Control
Wireless Compromise						Role Identification		Modify Alarm Settings	Spoof Reporting Message	Manipulation of View
						Screen Capture		Modify Control Logic	Unauthorized Command Message	Theft of Operational Information
								Program Download		
								Rootkit		
								System Firmware		
								Utilize/Change Operating Mode		

ATT&CK for Enterprise
ATT&CK for ICSs

Choi et al., "Probabilistic Attack Sequence Generation and Execution Based on MITRE ATT&CK for ICS Datasets", in Proc. of ACM CSET, 2021

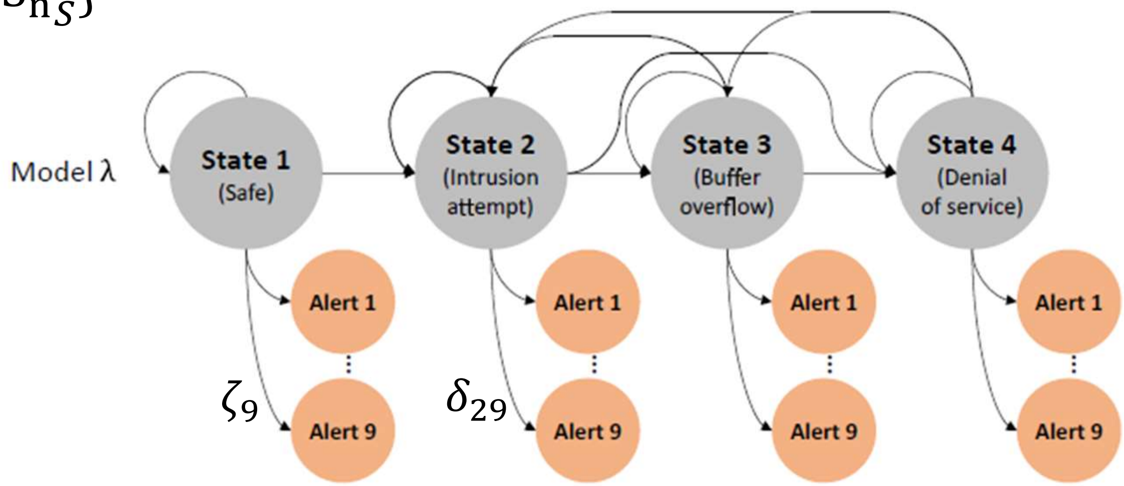
Attack and Observation Model

- Set of attacker states $\mathcal{S} = \{s_1, \dots, S_{n_\mathcal{S}}\}$
- State at time t: S_t
- Set of alerts $\mathcal{J} \ = \{1, \dots, J\}$
- True alert probability

$$\delta_{ij} = P(Y_t^j = 1 | S_t = s_i)$$

- False alert probability

$$\zeta_j = P(Y_t^j = 1 | S_t = s_1)$$



Defender Model

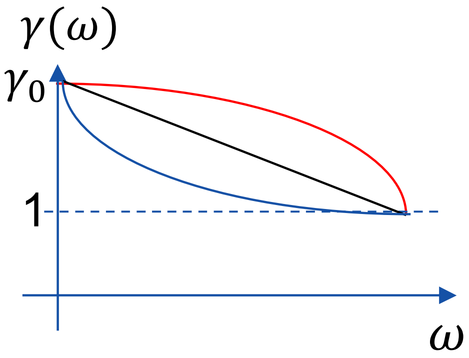
- **Observes** alerts Y_t at time t
- **Investigates** up to I alerts $v \subseteq Y_{1:t}$
  - Investigation outcome o_t
- **Human model:** Investigation error probability ω
- Confidence function

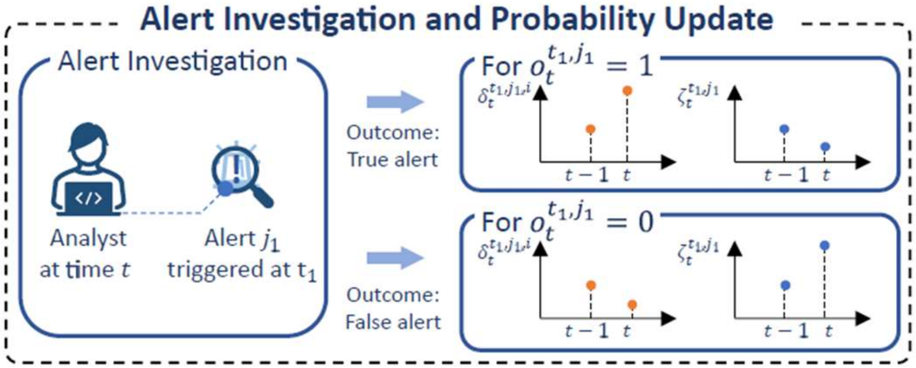
$$\gamma(\omega) = \begin{cases} 2(1-\gamma_0)\omega + \gamma_0, & (linear), \\ 4(1-\gamma_0)\omega^2 + \gamma_0, & (concave), \\ 4(\gamma_0 - 1)(\omega - 0.5)^2 + 1 & (convex), \end{cases}$$

- Update of HMM Observation Model

$$\delta_t^{t',j,i} = \begin{cases} \frac{1}{\gamma(\omega)} \delta_{t-1}^{t',j,i} & if \;\; o_t^{t',j} = 0 \\ min\left(\gamma(\omega)\delta_{t-1}^{t',j,i}, 1\right) & if \;\; o_t^{t',j} = 1 \end{cases}$$

$$\zeta_t^{t',j} = \begin{cases} min\left(\gamma(\omega)\zeta_{t-1}^{t',j}, 1\right) & if \;\; o_t^{t',j} = 0 \\ \frac{1}{\gamma(\omega)} \zeta_{t-1}^{t',j} & if \;\; o_t^{t',j} = 1 \end{cases}$$





Kim et al, "Human-in-the-loop Cyber Intrusion Detection Using Active Learning" IEEE TIFS, 2024

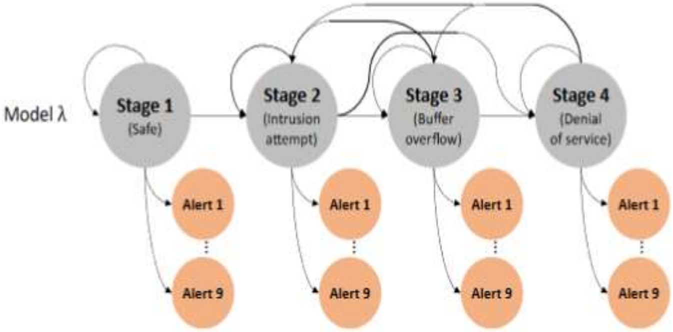
Defender's Problem

- **Defender objective:** Minimize mean time to detection

$$\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \sup_{t_{1\to 2}>0} \mathbb{E}^{(t_{1\to 2})}[d^\kappa - t_{1\to 2}]$$

  - **Subject to:** Constraint on false positive rate

$$\mathbb{E}^{(\infty)}[d^\kappa] \geq \tau.$$

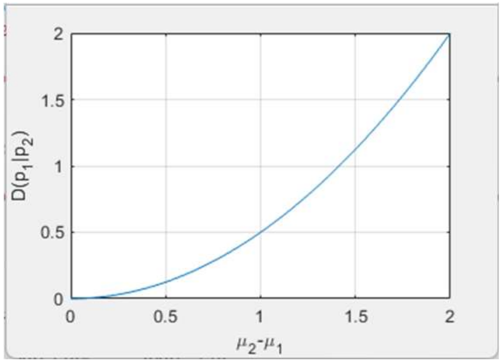


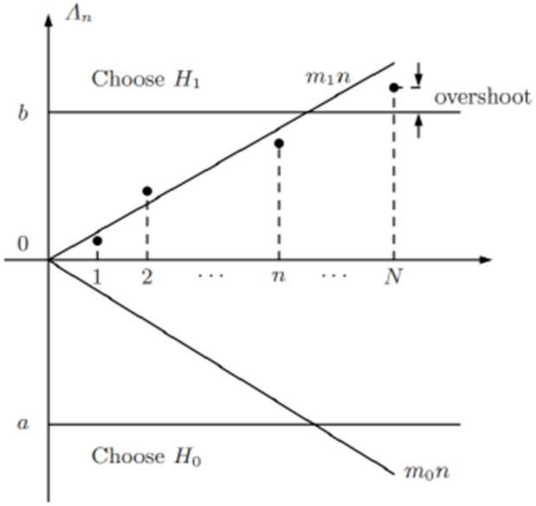
Kim et al, "Human-in-the-loop Cyber Intrusion Detection Using Active Learning" IEEE TIFS, 2024

Background: Sequential Hypothesis Testing

- Generalized likelihood ratio test
  - Composite hypothesis $\mathcal{H}_1 = \{h_1, \dots, h_H\}$
  - Detection rule $\eta_t = \begin{cases} \mathcal{H}_1 \; if \; \dfrac{\max\limits_{h \in \mathcal{H}_1} P(\mathrm{Y}|h)}{P(\mathrm{Y}|h_0\,)} > \Theta \\ \quad \text{otherwise} \end{cases}$

- Asymptotic behavior
  - Risk
    - $R_h \triangleq \max\limits_{h' \neq h} P_{h'}(\eta_t = h)$
  - Expected detection time
    - $E\left[t_d^h\right] \geq \dfrac{-logR_h}{D(p_h||p_{h'})}(1 + o(1))$





B. C. Levy, Principles of signal detection and parameter estimation. Springer, 2008.

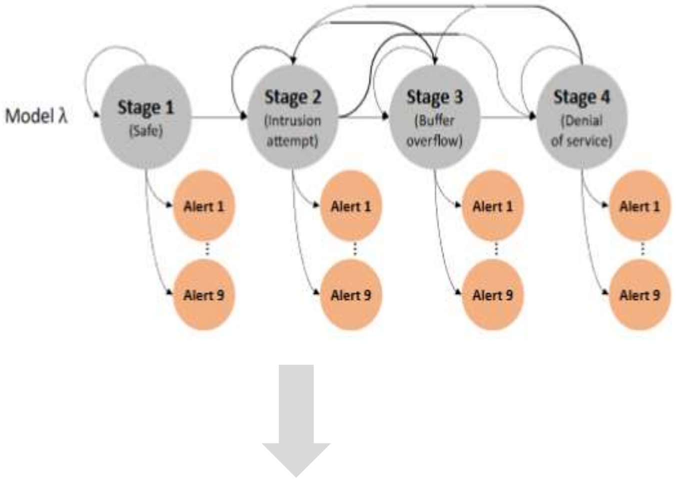
Defender’s Problem

- **Defender objective:** Minimize mean time to detection

$$\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \sup_{t_{1\to 2}>0} \mathbb{E}^{(t_{1\to 2})}[d^\kappa - t_{1\to 2}]$$

  - **Subject to:** Constraint on false positive rate

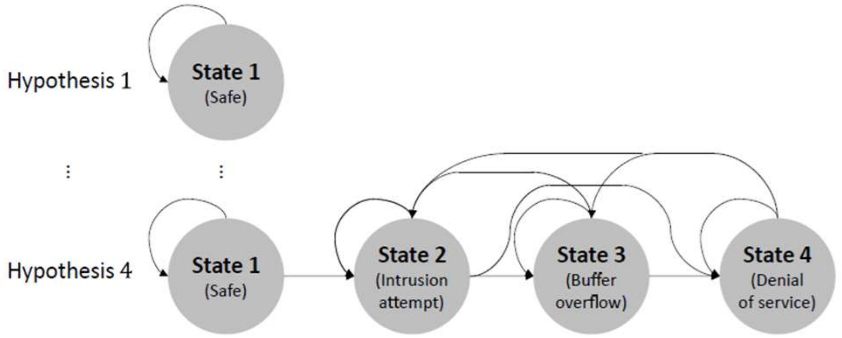
$$\mathbb{E}^{(\infty)}[d^\kappa] \geq \tau$$



- **Generating Alternative Hypotheses**
  - Most likely hypothesis at time t

$$\hat{h} = argmax_{h \in \mathcal{H}} P_h(Y_{1:t} \,|\mathcal{F}_t, v_t^\kappa)$$

  - Likelihood ratio
    > $S_t^\kappa = \frac{P_{\hat{h}}(Y_t|\mathcal{F}_t, v_t^\kappa)}{P_1(Y_t|\mathcal{F}_t, v_t^\kappa)}$



Active Learning for Quickest Detection

- **Optimal detection rule without active learning**
  - Generalized likelihood ratio test
- **Two candidate policies**
  - Max-ratio policy

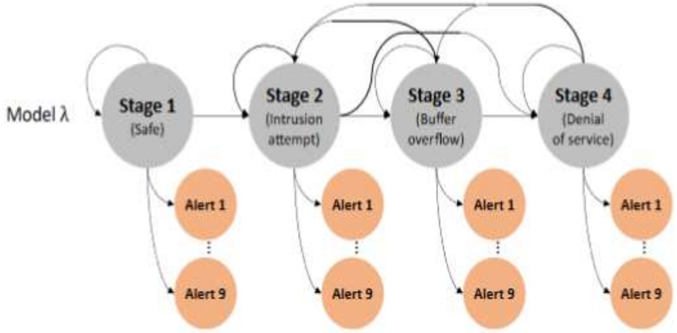
    → Set of alerts that maximizes the expected probability ratio

$$\mathcal{V}_t^{MR} = \underset{v_t \subseteq Y^+_{1:t}, |v_t| \le B}{\arg\max} \left| \mathbb{E}\left[\frac{p_{\hat{h}}(Y_{1:t} = y_{1:t} | \mathcal{F}_t = f_t, \mathcal{V}_t = v_t)}{p_1(Y_{1:t} = y_{1:t} | \mathcal{F}_t = f_t, \mathcal{V}_t = v_t)} \right] \right|$$

  - Max KL Divergence

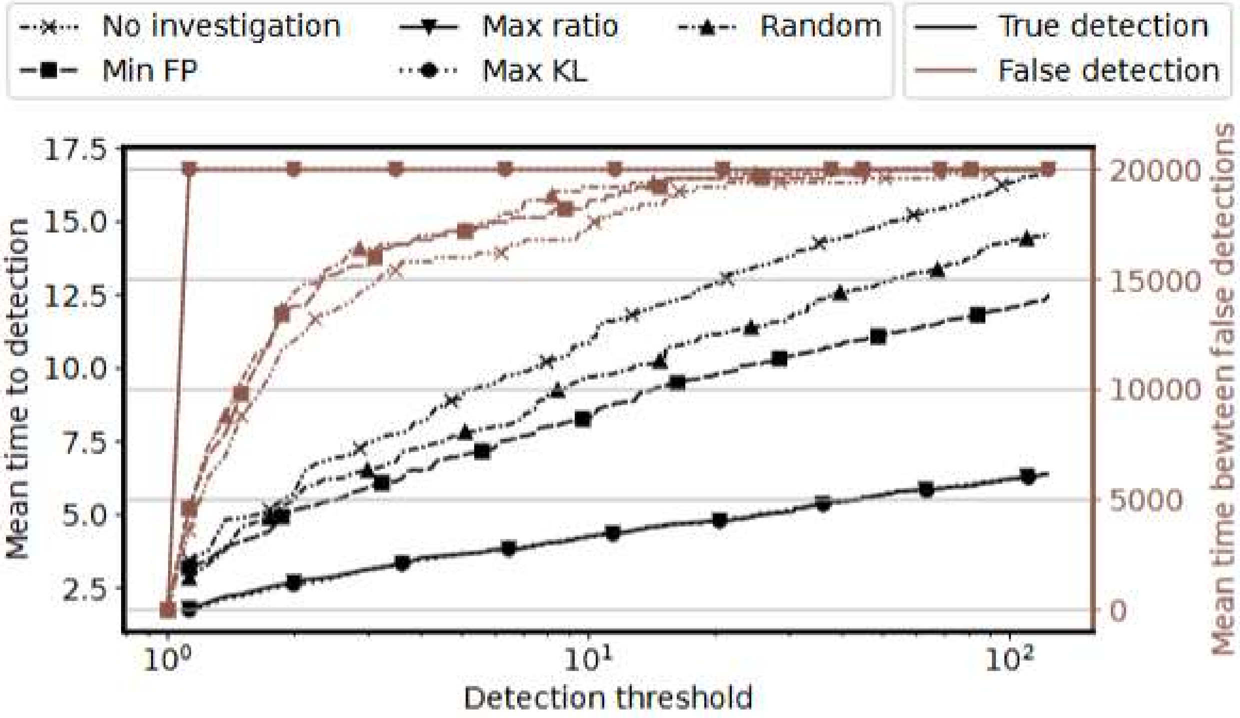
    → Set of alerts that maximize the KL divergence of the distribution of observed alerts after investigation

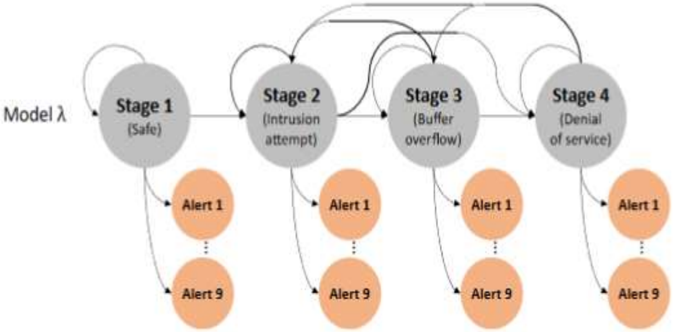
$$\mathcal{V}_t^{MKL} = \underset{v_t \subseteq Y^+_{1:t}, |v_t| \le B}{\arg\max} \mathbb{E}\Big[\sum_{t'=1}^{t} D\big(\mathbb{P}_{\hat{h}}(Y_{t'} = y_{t'} | \mathcal{F}_t = f_t, \mathcal{V}_t = v_t) \parallel \mathbb{P}_1(Y_{t'} = y_{t'} | \mathcal{F}_t = f_t, \mathcal{V}_t = v_t)\big)\Big]$$
$$- \sum_{t'=1}^{t} D(\mathbb{P}_{\hat{h}}(Y_{t'} = y_{t'} | \mathcal{F}_t = f_t) \| \mathbb{P}_1(Y_{t'} = y_{t'} | \mathcal{F}_t = f_t))$$

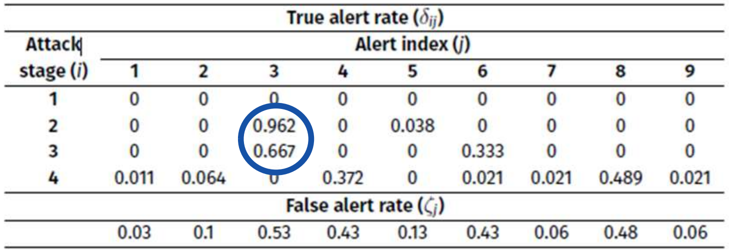


Kim et al, "Human-in-the-loop Cyber Intrusion Detection Using Active Learning" IEEE TIFS, 2024

Detection Performance

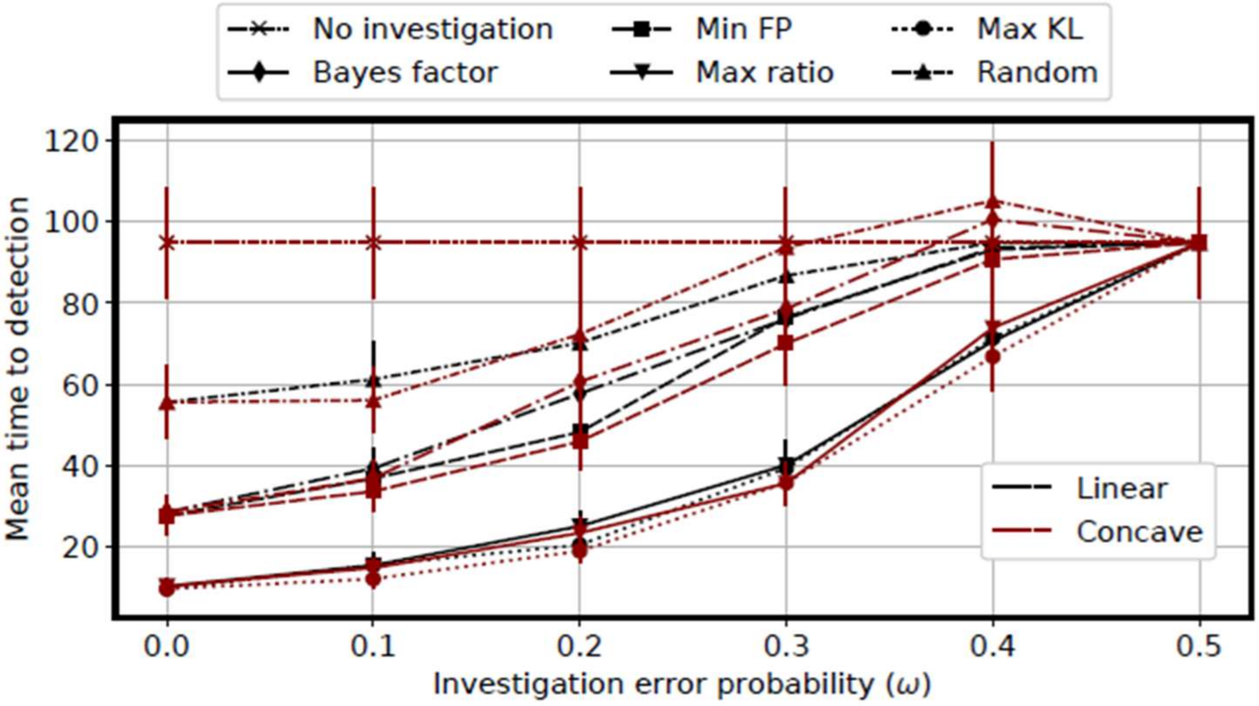




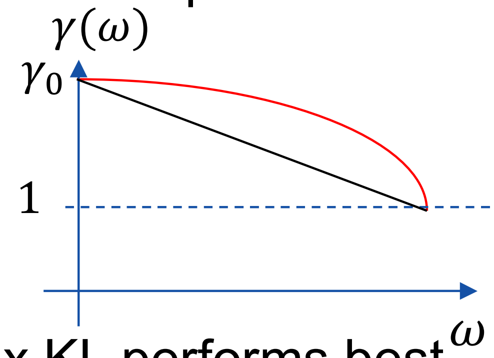


	True alert rate (δ_{ij})								
Attack	Alert index (j)								
stage (i)	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0.962	0	0.038	0	0	0	0
3	0	0	0.667	0	0	0.333	0	0	0
4	0.011	0.064	0	0.372	0	0.021	0.021	0.489	0.021
	False alert rate (ζ_j)								
	0.03	0.1	0.53	0.43	0.13	0.43	0.06	0.48	0.06

Impact of the Human Model

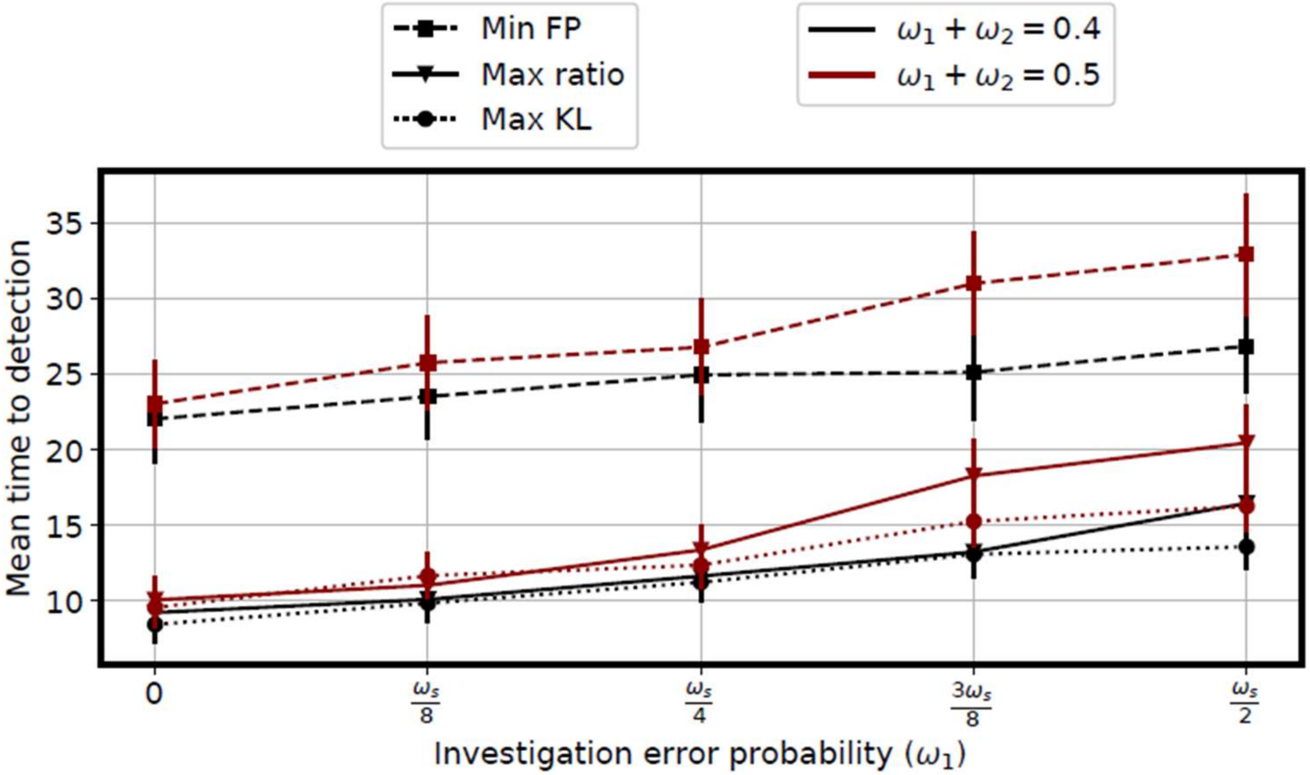


- Concave confidence function superior



- Max KL performs best

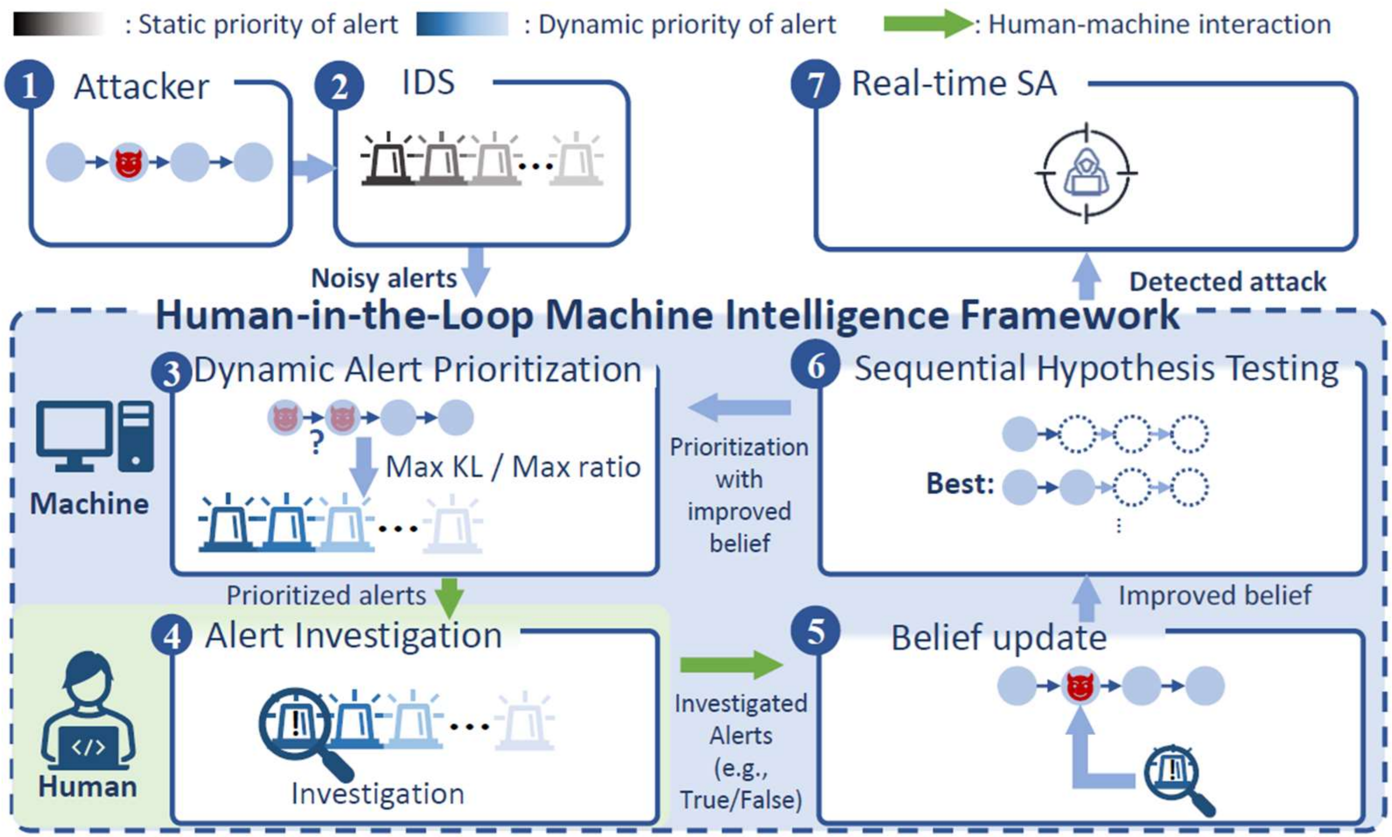
Expertise is Important



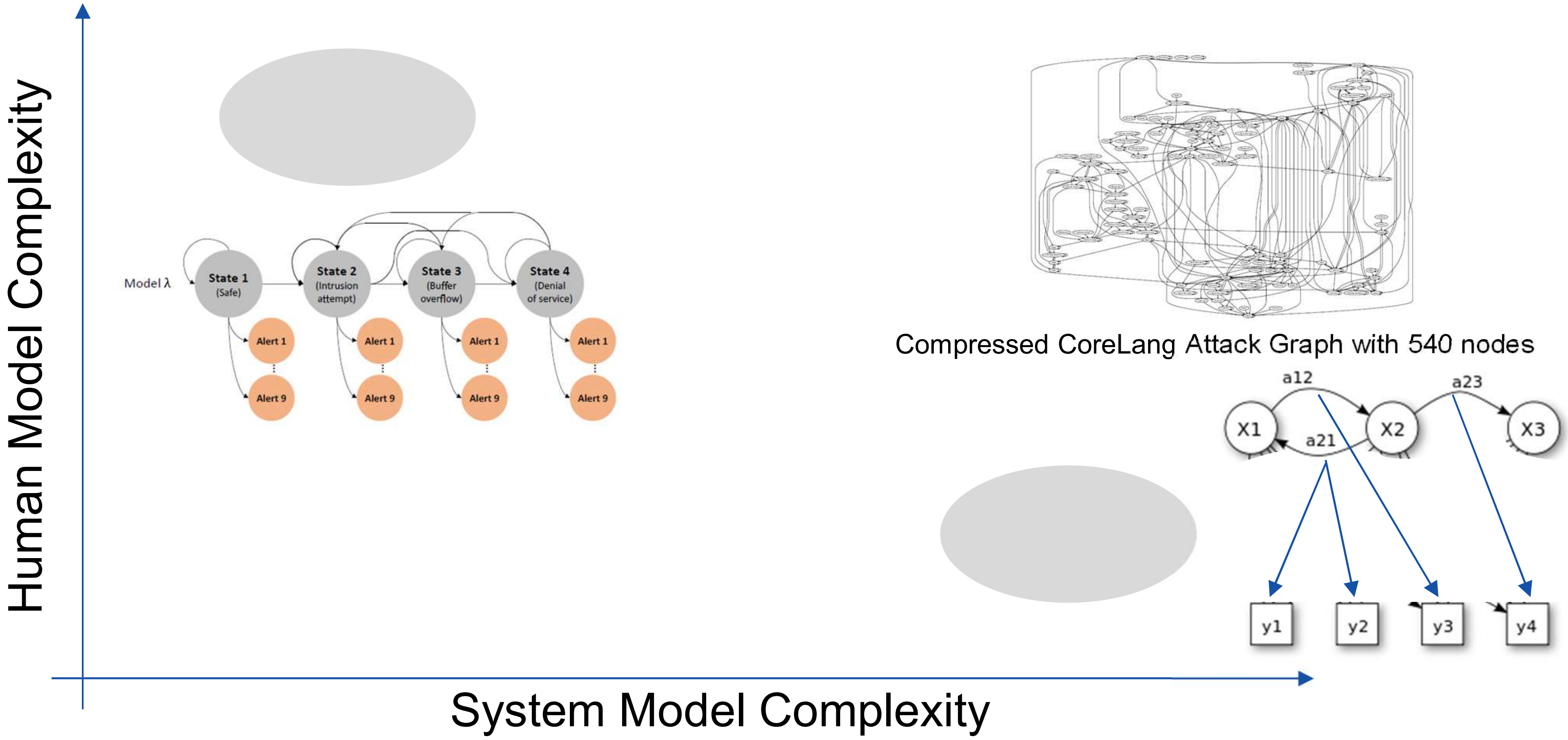
- 2 experts with potentially varying expertise
- Heterogenous expertise is preferrable

Kim et al, “Human-in-the-loop Cyber Intrusion Detection Using Active Learning" IEEE TIFS, 2024

Human-in-the-Loop AI Framework Revisited


: Static priority of alert
: Dynamic priority of alert
: Human-machine interaction
1 Attacker
2 IDS
7 Real-time SA
Noisy alerts
Detected attack
Human-in-the-Loop Machine Intelligence Framework
3 Dynamic Alert Prioritization
?
Max KL / Max ratio
Machine
6 Sequential Hypothesis Testing
Prioritization with improved belief
Best:
Prioritized alerts
Improved belief
4 Alert Investigation
Investigation
Human
Investigated Alerts (e.g., True/False)
5 Belief update

Framework Design Space

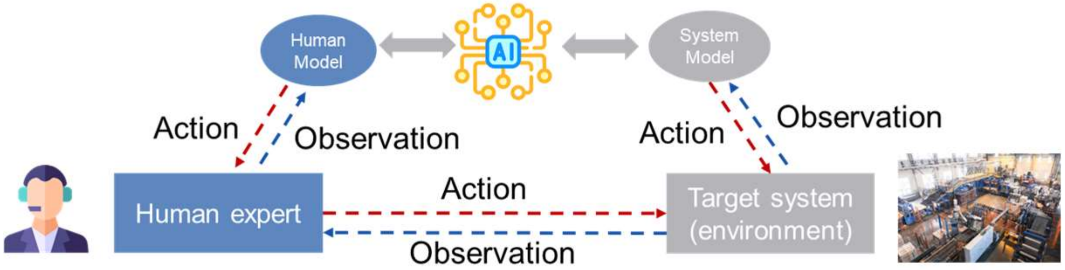


Compressed CoreLang Attack Graph with 540 nodes

Katsikeas et al. “An attack simulation language for the IT domain,” in *Proc. of Int. Workshop on Graphical Models for Security*, pp. 67–86, 2020.

Conclusion

- **Human-in-the-loop-AI for cyber resilience**
  - Efficient use of human resources and ML
  - Human skills and behavior vs. system model complexity
  - Improved accuracy and lower time to detection
- **Many open questions**
  - How to model human behaviour
    - > *Trust, psychological aspects*
    - > Affects the design of AI algorithms
  - How to apply the concept to CPS
  - Vulnerability to an adaptive adversary in a game theoretical framework
  - Integration with threat hunting
  - Semi-autonomous incident response



KTH ROYAL INSTITUTE OF TECHNOLOGY

Boosting Cyber Resilience with Human-in-the-loop AI

György Dán

IEEE CNS 2024 Workshop on Cyber Resilience

C3.ai Digital Transformation Institute

digital futures