First Person Attacker Simulation in MAL

CDIS Spring Conference, 22 May 2025

Sandor Berglund (sandorb@kth.se) Mathias Ekstedt (mekstedt@kth.se)

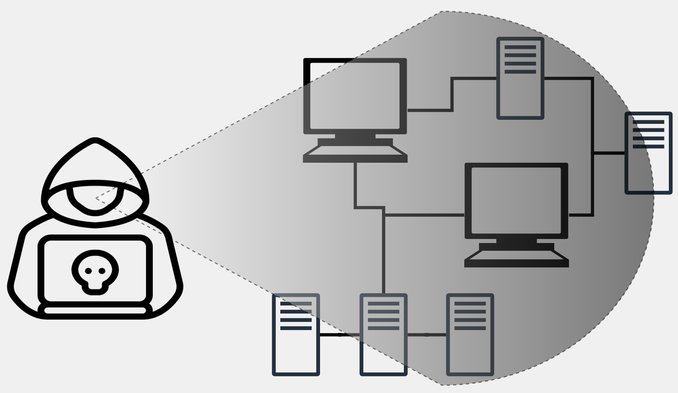
Centre for Cyber Defense and Information Security (CDIS)
KTH Royal Institute of Technology

Motivation

What do hackers do under an ongoing cyber attack?

- Malicious organizations seldom share information
- Emulating "real hackers" is costly
- Attacks can come from unknown entities

To resolve this and answer the question, a simulation of the cyber attack is employed. The simulation is done from the attacker's perspective.

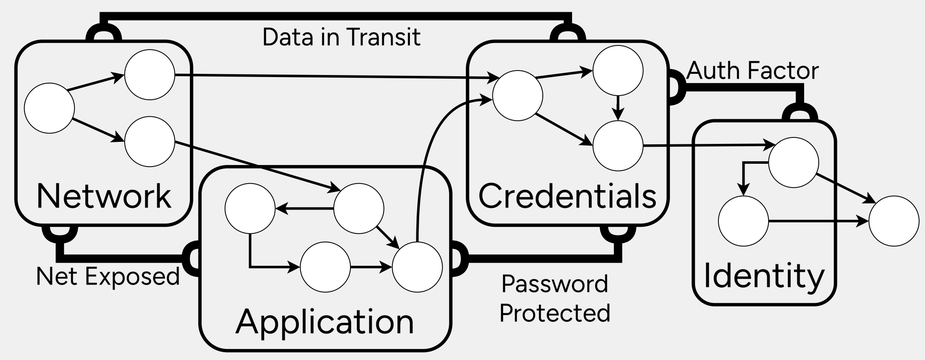


MAL Attack Graphs

Attack graphs according to the **Meta Attack Language** (**MAL**) specification is used for the attack simulations. **MAL** specifies:

- Classes of entities that can exist in the system.
- Attack steps that can be performed on specific entities.
- Types of relations between entities of certain classes.

By filling in the system specific of these definitions, an attack graph of what is possible for the attacker to do can be generated.



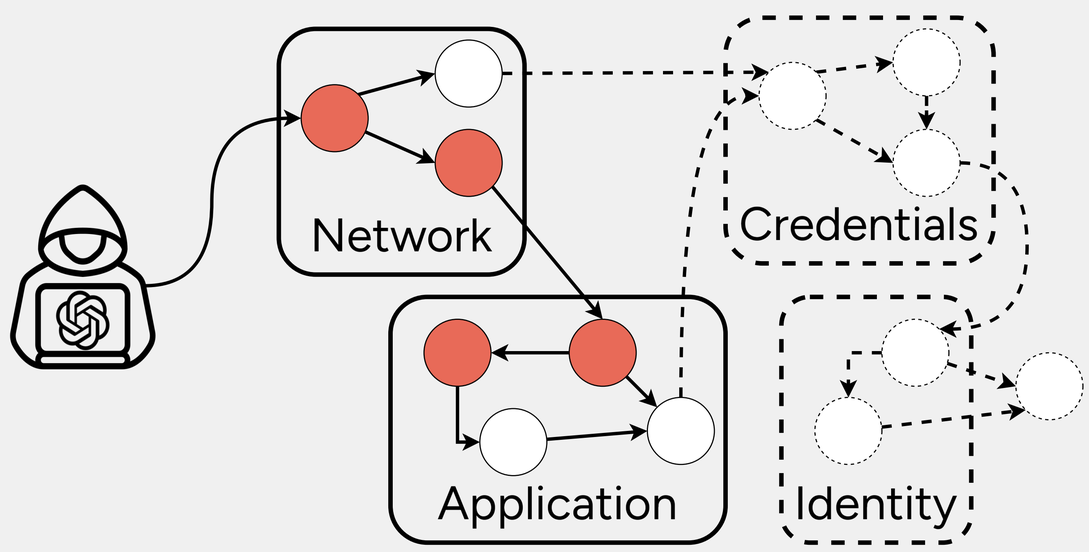
LLM as the Attacker

We plan to compare **Large Language Models** (**LLM**s) with **humans** in terms of attack step selection. The users will be able to do tactical simulations in a terminal interface.

```
0: LAN:connect                (Network)
> Enter action: 0
---
1: LAN:scanPorts              (Network)
2: LAN:eavesdrop              (Network)
> Enter action: 1
---
3: App17:connect              (Application)
> Enter action: 3
---
4: App17:scanPrivileges       (Application)
5: App17:readContainedData    (Application)
> Enter action: 4
---
6: App17:accessContainedData (Application)
> Enter action:
```

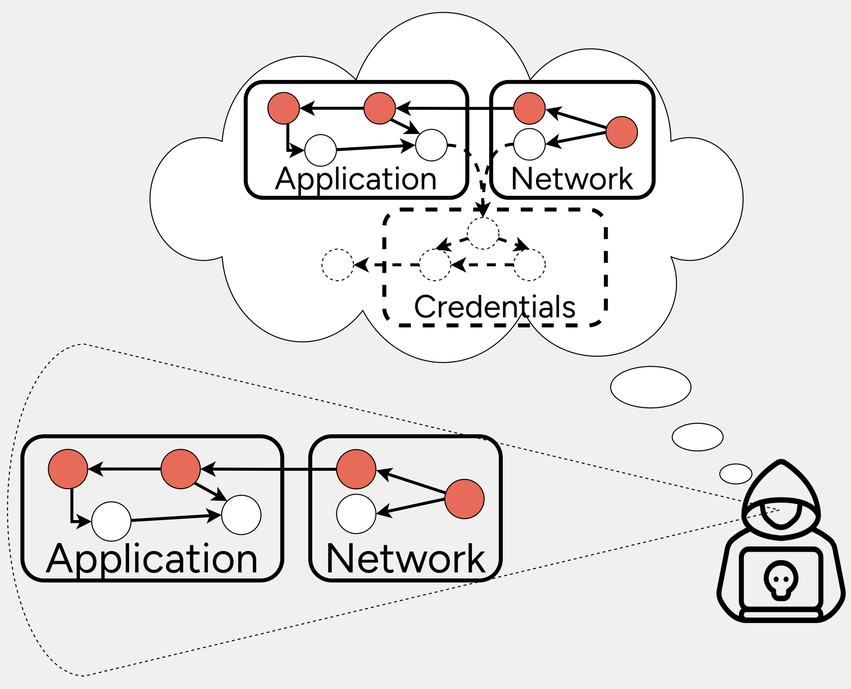
Users have to find out which attack steps are worthwhile and allow them to penetrate the system. The *fog of war* about what lies further in the system creates uncertainty and forces the users to explore.

Hypothesis: **LLM**s can reason about the tactics of a cyber attack like humans.



In collaboration with:
Simon Gökstorp (gokstorp@kth.se)
Pontus Johnson (pontusj@kth.se)

Bayesian Planning and Inference as the Attacker



Current research regards seeing how the Bayesian modeling area **Active Inference** can be leveraged to represent attackers.

True state:	$s \in \mathcal{S}$	Allowable actions:	$u \in \mathcal{A}$
Observation:	$o \in \mathcal{O}$	Policy:	$\pi \in \mathcal{A}^N$

- **Goal**: Preferred observations $p(o|C)$
- **World model**: Approximate posterior over states $q(s)$

Likelihood of preferred observations $p(o|C)$ is maximized through minimizing **Variational Free Energy** (**VFE**) and **Expected Free Energy** (**EFE**).

$$\text{VFE: } F_\pi = \underbrace{D_{KL}[q(s|\pi)\|p(s|\pi)]}_{\text{Complexity}} - \underbrace{\mathbb{E}_{q(s|\pi)}[\ln p(o|s)]}_{\text{Accuracy}}$$

$$\text{EFE: } G_\pi = -\underbrace{\mathbb{E}_{q(o,s|\pi)}[\ln q(s \mid o,\pi) - \ln q(s|\pi)]}_{\text{Epistemic value}} - \underbrace{\mathbb{E}_{q(o|\pi)}[\ln p(o|C)]}_{\text{Pragmatic value}}$$