Next Steps in Cyber Blue Team Automation — Leveraging the Power of LLMs

Allard Dijk[1], Roland Meier[2], Cosimo Melella[3],
Mauno Pihelgas[4], Risto Vaarandi[4], Vincent Lenders[2]

28.05.2025

CYCON

(1) NL Defence Academy
NLDA

(2) Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
armasuisse

(3) CCDCOE
NATO COOPERATIVE
CYBER DEFENCE
CENTRE OF EXCELLENCE

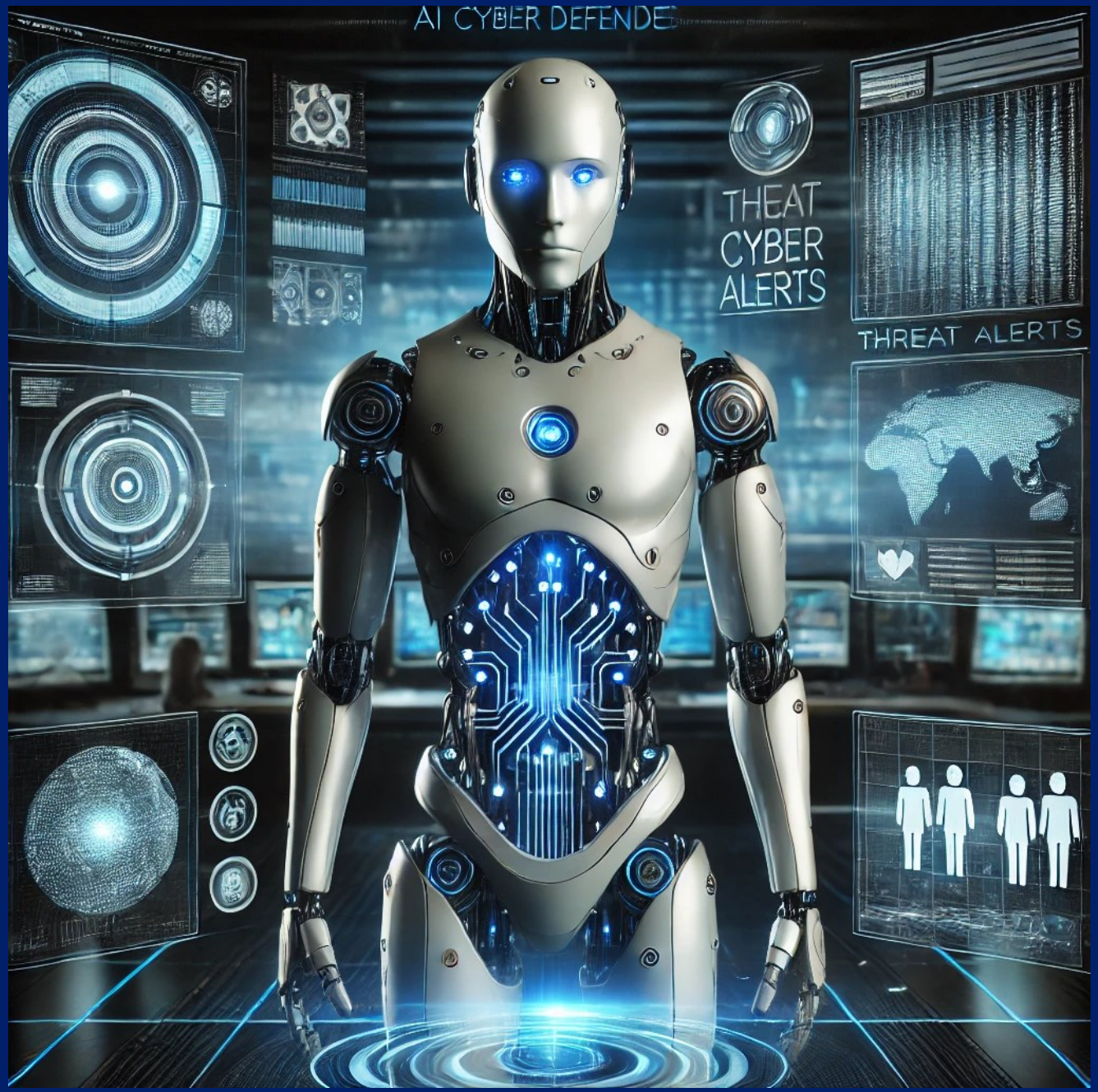
(4) TAL TECH

CES Attend Exhibit Plan Your Visit Explore CES Discover SIGN UP APPLY TO EXHIBIT

CES INNOVATION AWARDS® / 2025 /

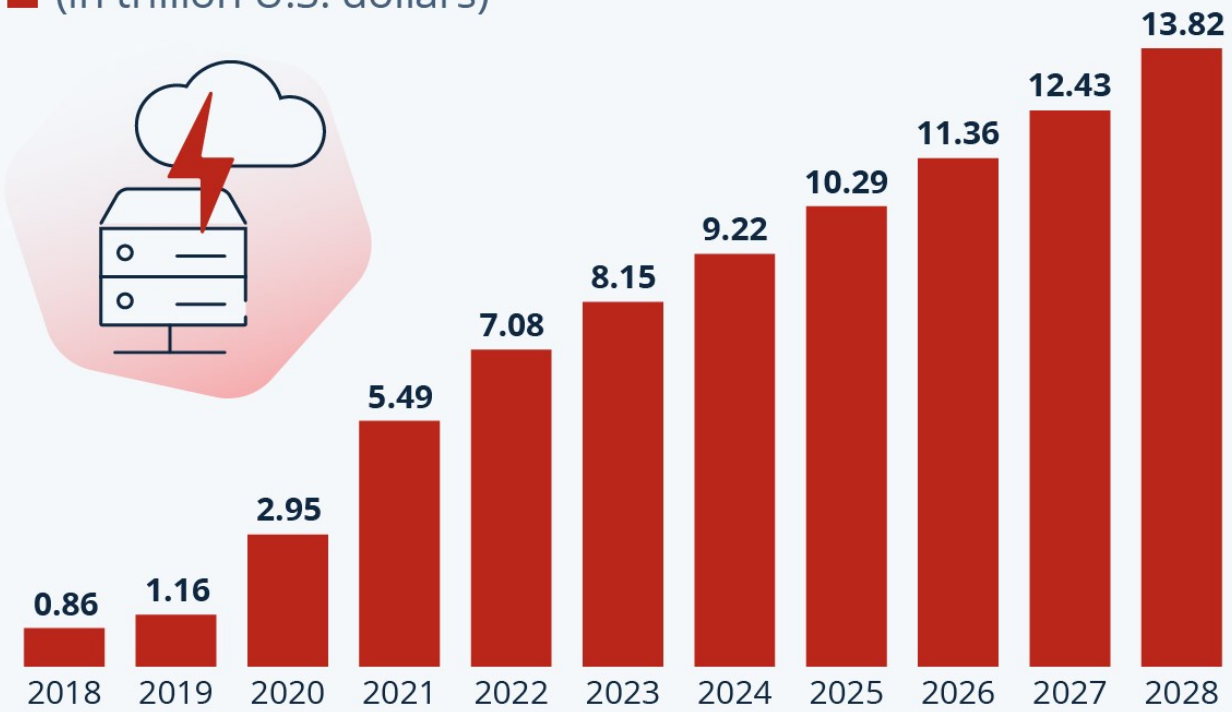
CES Innovation Awards 2025 Honoree

4-Door Refrigerator with AI Home and AI Vision Inside 2.0


AI CYBER DEFENDE
THEAT
CYBER
ALERTS
THREAT ALERTS

Cybercrime Expected To Skyrocket

Estimated annual cost of cybercrime worldwide (in trillion U.S. dollars)



As of Sep. 2023. Data shown is using current exchange rates.

Source: Statista Market Insights

Forbes

FORBES > LEADERSHIP > CAREERS

Nearly 4 Million Cybersecurity Jobs Are Vacant: Here's Why You Should Consider Breaking Into This Sector

Jack Kelly Senior Contributor ⓘ

Jack Kelly covers career growth, job market and workplace trends.

Follow

Aug 16, 2024, 06:00am EDT

Cybersecurity consistently ranks among the top areas for job growth and demand within the broader ... [+] GETTY

We developed a framework for an automated Blue Team a couple of years ago

2021 13th International Conference on Cyber Conflict
Going Viral
T. Jančárková, L. Lindström, G. Visky, P. Zotz (Eds.)

Towards an AI-powered Player in Cyber Defence Exercises

Roland Meier
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

Artūrs Lavrenovs
NATO CCDCOE
Tallinn, Estonia
arturs.lavrenovs@ccdcoe.org

Kimmo Heinäaro
NATO CCDCOE
Tallinn, Estonia
kimmo.heinaaro@mil.fi

Luca Gambazzi
Science and Technology
armasuisse
Thun, Switzerland
luca.gambazzi@armasuisse.ch

Vincent Lenders
Science and Technology

CYCON

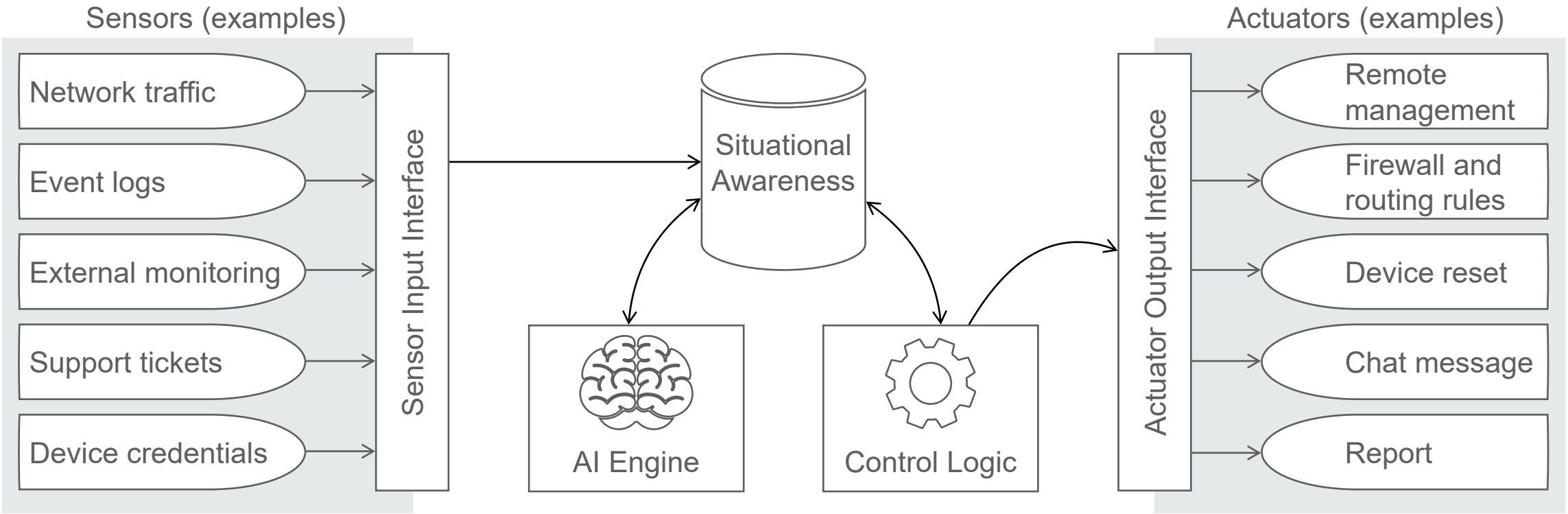
How can automation / AI help for cyber defense?

And eventually…

What would it take to have a fully automated Blue Team in a future iteration of Locked Shields?

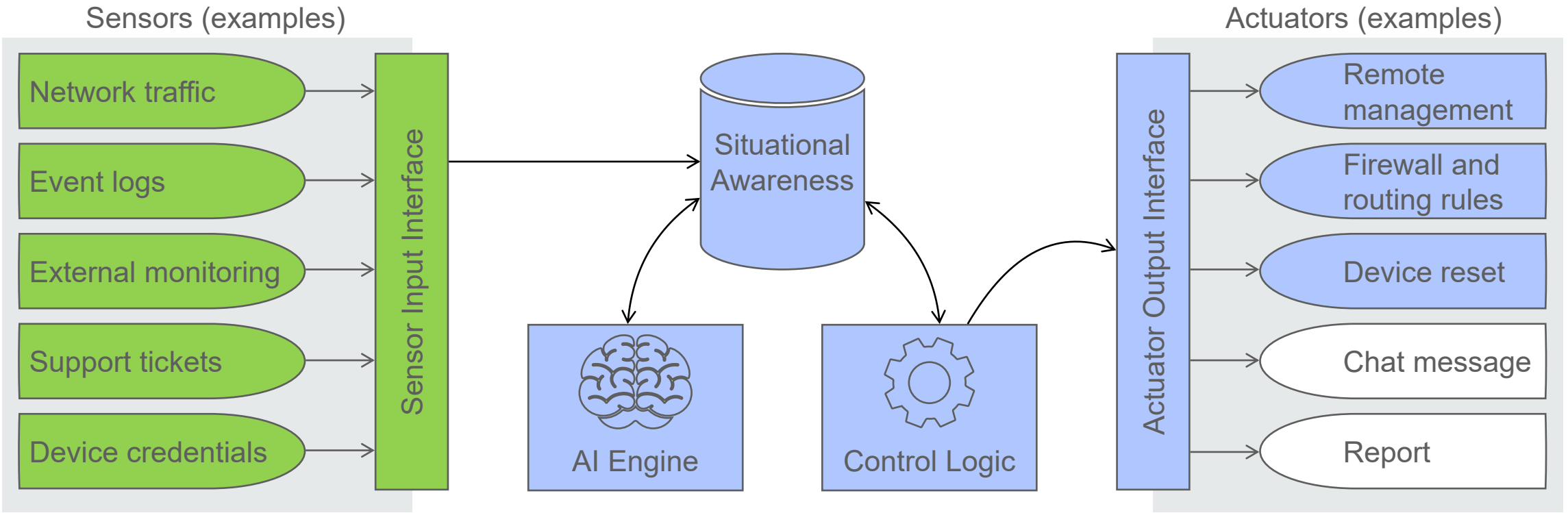
CYCON

Automated Blue Team framework overview



CYCON

In the meantime, the framework is implemented and we continuously extend it



CYCON

Locked Shields is the largest live-fire global cyber defense exercise

Picture: NATO CCDCOE

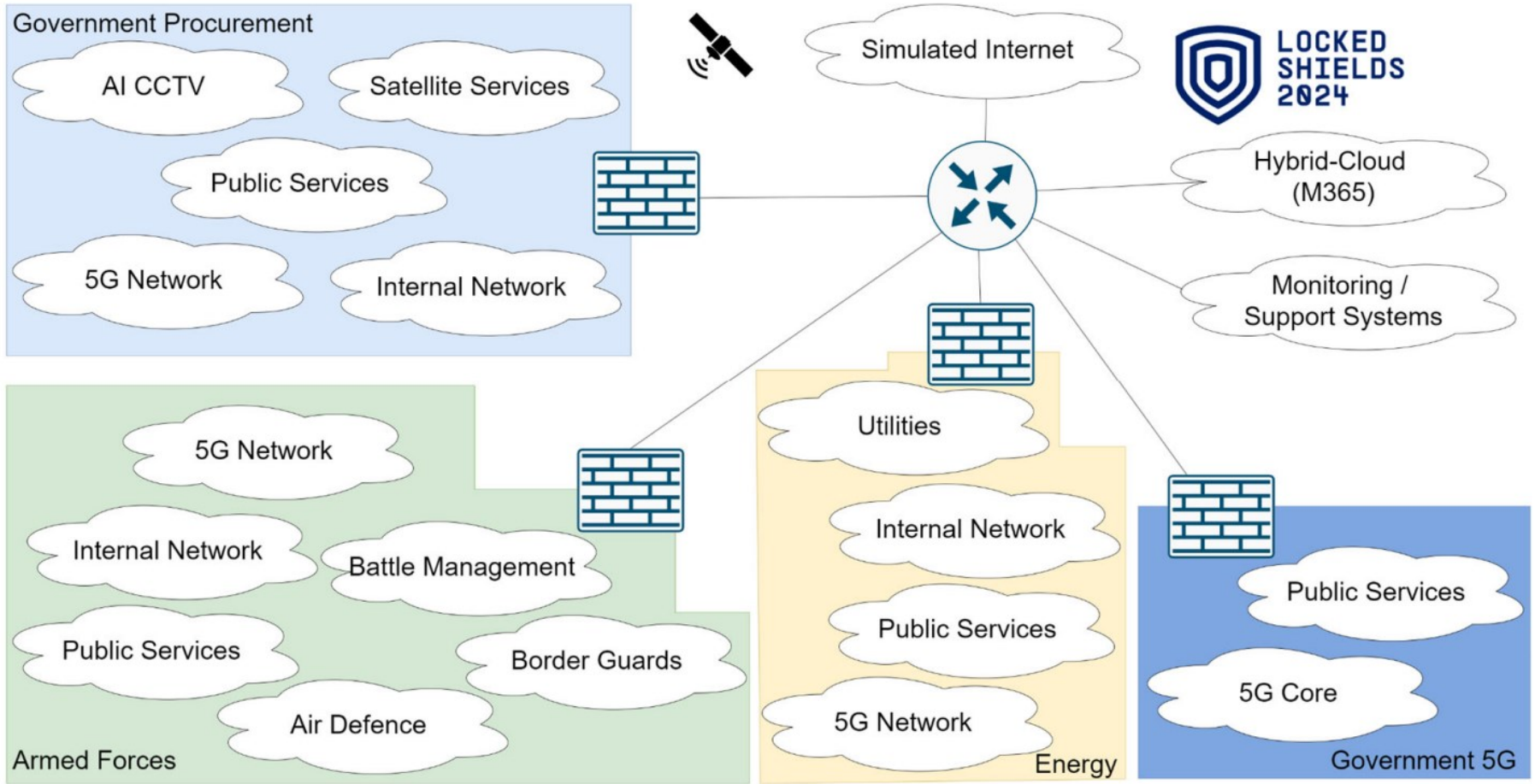
Locked Shields is the largest live-fire global cyber defense exercise

- Red Team vs. Blue Team exercise

Attackers	Defenders
1 Team	~1 Team / country

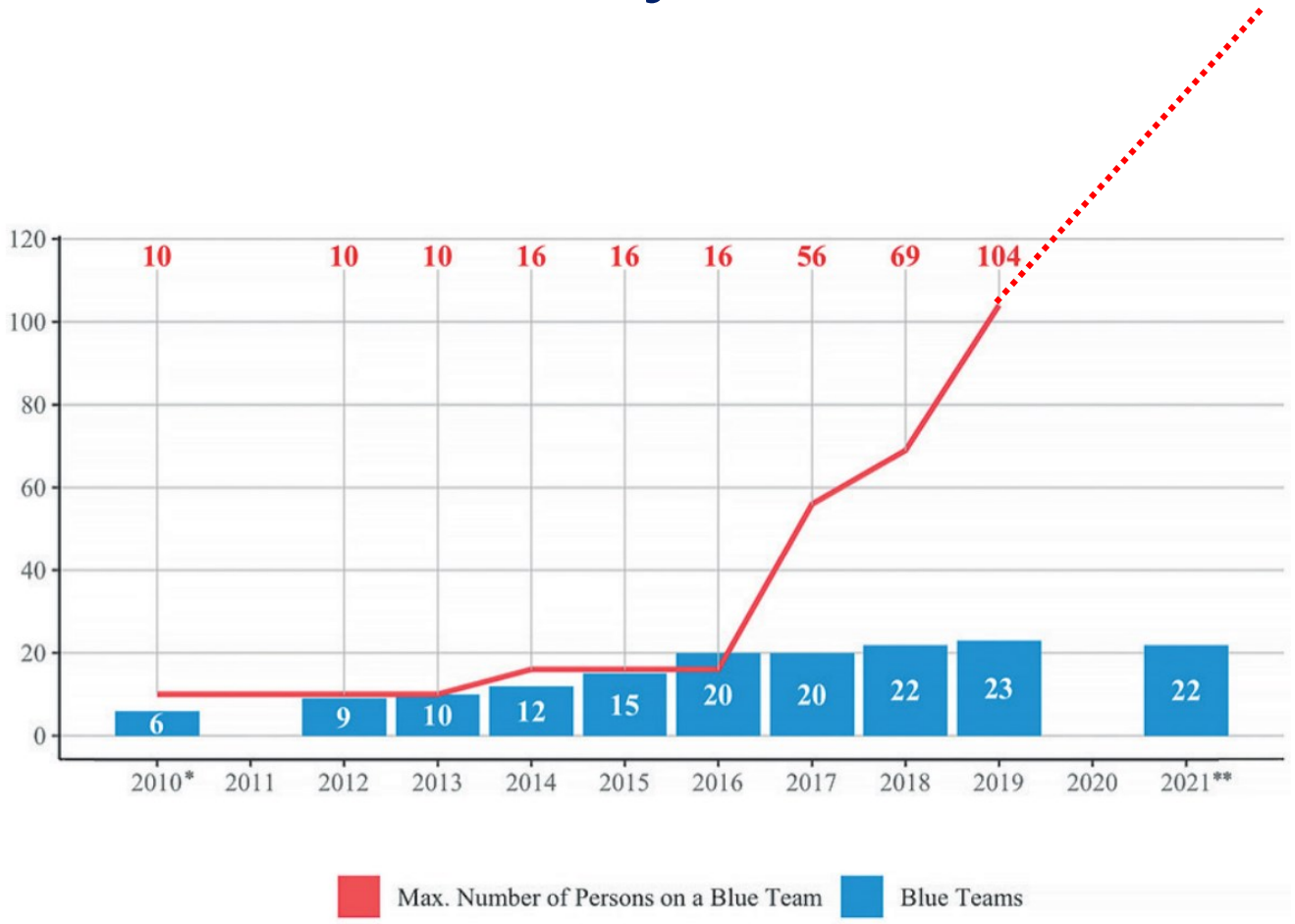
Picture: NATO CCDCOE

Locked Shields is the ideal testing ground for AI research



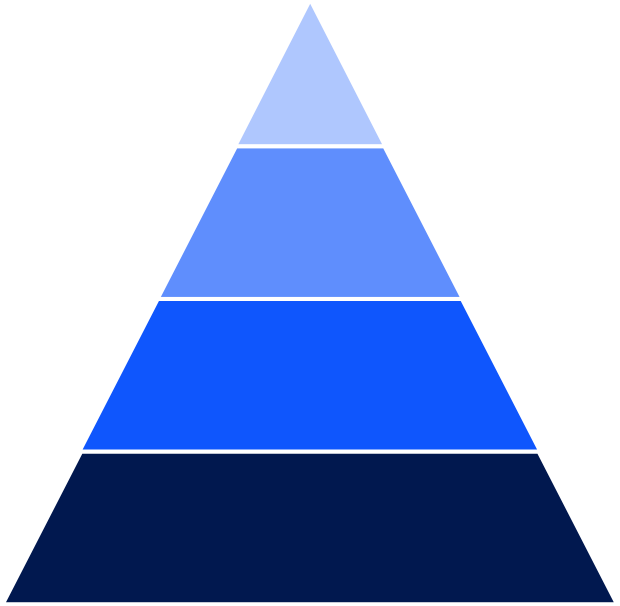
CYCON

The number of people required in a Blue Team continuously increases



[Smeets, Max. "The role of military cyber exercises: A case study of Locked Shields." *CyCon 2022*]

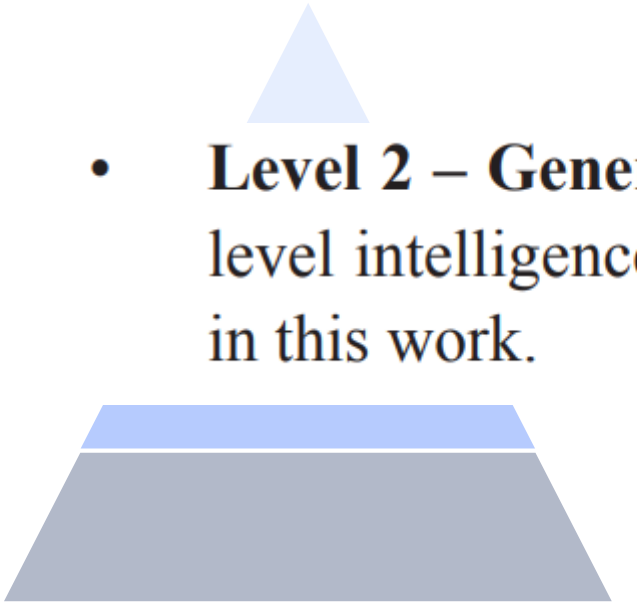
We distinguish four levels of AI



Level 3: Super AI	Surpasses human intelligence
Level 2: General AI	Mimics human intelligence
Level 1: Limited-memory narrow AI	“Machine learning”
Level 0: Reactive narrow AI	Signature-based decisions

CYCON

We distinguish four levels of AI



Level 3: Super AI

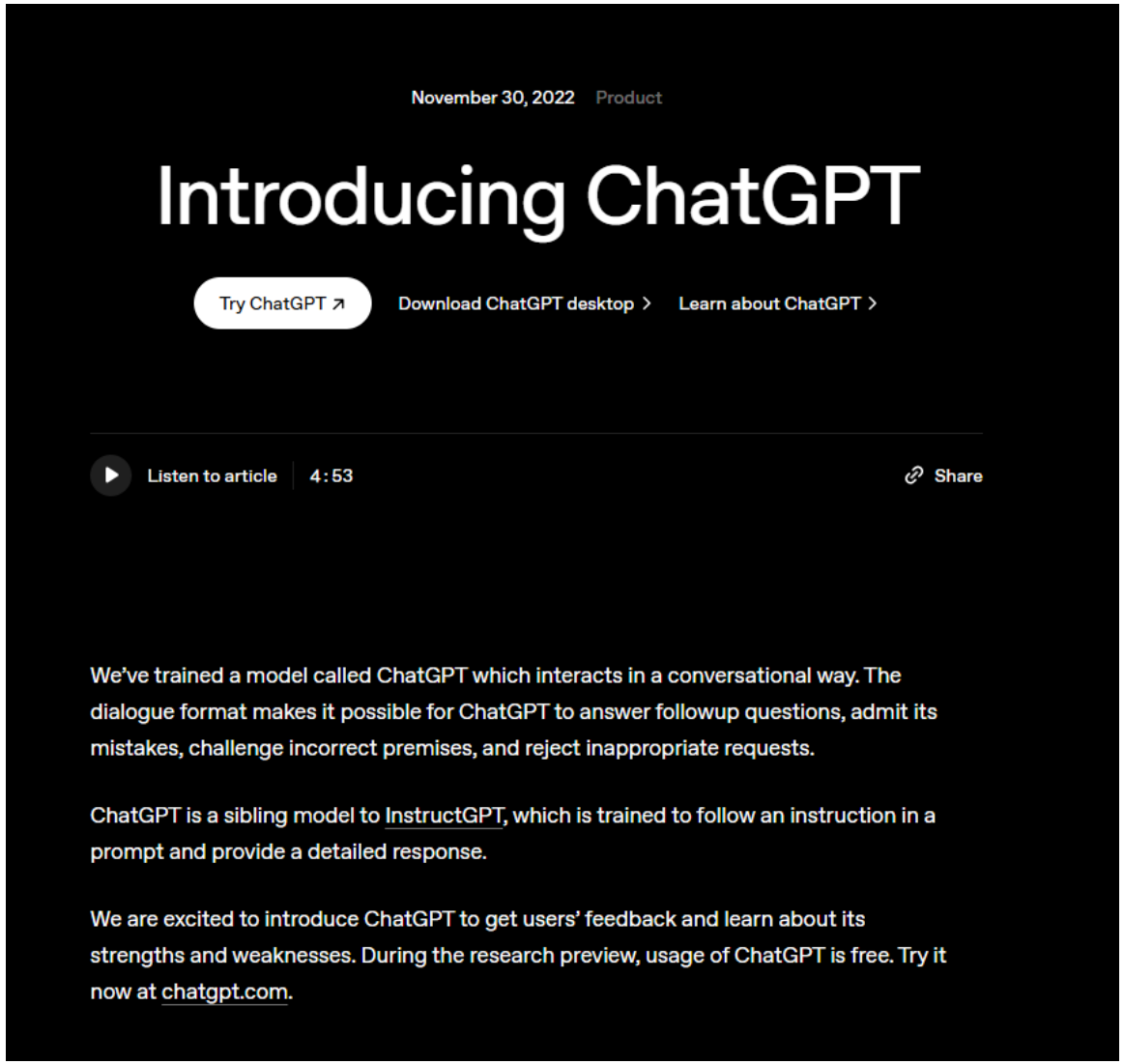
Surpasses human intelligence

- **Level 2 – General AI / Strong AI / Deep AI:** Level-2 AI mimics human-level intelligence but it does not exist yet. We, therefore, do not consider it in this work.

Level 0: Reactive narrow AI

Signature-based decisions

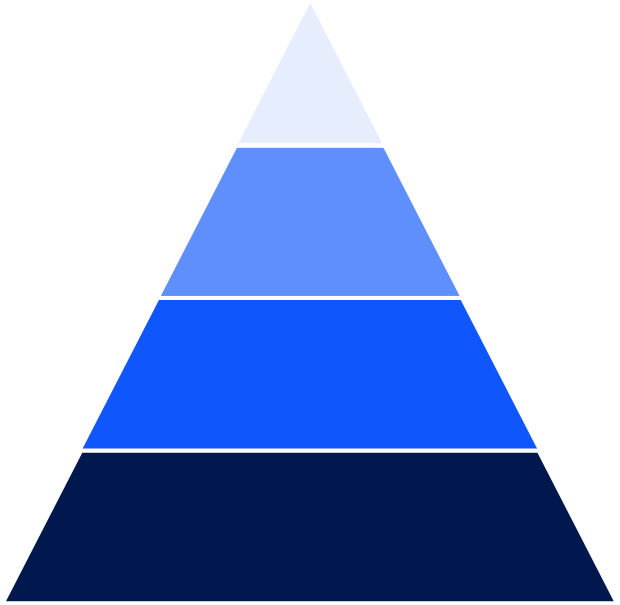
CYCON


November 30, 2022 Product
Introducing ChatGPT
Try ChatGPT ↗
Download ChatGPT desktop >
Learn about ChatGPT >
Listen to article 4:53
Share
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.
ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.
We are excited to introduce ChatGPT to get users' feedback and learn about its strengths and weaknesses. During the research preview, usage of ChatGPT is free. Try it now at chatgpt.com.

CYCON

~~Only level 0 and level 1 exist today~~

Levels 0 – 2 exist today



Level 3: Super AI	Surpasses human intelligence
Level 2: General AI	Mimics human intelligence
Level 1: Limited-memory narrow AI	“Machine learning”
Level 0: Reactive narrow AI	Signature-based decisions

CYCON

Generative AI offers many new opportunities

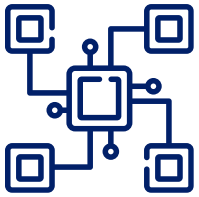
Automated patching

Ticket parsing

Network analysis

Report generation

Log analysis



Data integration

CYCON

Generative AI can detect and fix vulnerabilities and misconfigurations

Problem

- Misconfigurations in systems remain a critical security risk
- Manual analysis is time-consuming and error-prone

Potential for generative AI

- Analyze large codebases and configurations of diverse applications

Relevance in Locked Shields

- Blue Teams must harden many systems quickly

Existing approaches

- LLM-driven vulnerability detection (e.g., LProtector)
- Code-generation tools (e.g., GitHub Copilot)

CYCON

Generative AI can analyze network traffic and detect malicious activities

Problem

- Real-time identification of malicious activities in massive traffic volumes is difficult
- Encryption prevents analysis of contents

Relevance in Locked Shields

- Large variety of systems, applications and protocols requires a flexible detection approach

Potential for generative AI

- LLMs can interpret unstructured traffic data
- Transformer models detect malware patterns even with limited data

Existing approaches

- Interpreting network protocols similar to natural language
- RTIDS (transformer-based IDS)

CYCON

Generative AI can analyze event logs and extract relevant information

Problem

- Parsing logs manually is cumbersome
- Log volume is large

Potential for generative AI

- LLMs infer patterns with few or no examples
- LLMs handle variable and incomplete logs better than static rules

Relevance in Locked Shields

- Large variety of systems, applications and protocols requires a flexible log analysis approach

Existing approaches

- Automated template extraction
- Summarization of large log collections

CYCON

Given the above example, find the template for the following messages:

sshd[14331]: Failed none for alice from 10.1.3.3 port 28711 ssh2
sshd[31830]: Failed password for bob from 10.2.5.2 port 26926 ssh2

Answer directly with the template and explain the placeholders.

Template:

```
sshd[<*>]: Failed <*> for <*> from <*> port <*> ssh2
```

Explanation of placeholders:

- `<*>` (first): Represents the process ID, which varies in each log message.
- `<*>` (second): Represents the authentication method (e.g., "none", "password").
- `<*>` (third): Represents the username attempting to authenticate.

CYCON

Generative AI can interact with humans through natural language

Problem

- Support tickets and user reports require triage and technical translation

Potential for generative AI

- LLMs can parse natural language, extract key info, and generate actions and responses
- LLMs classify, prioritize, and summarize issues

Relevance in Locked Shields

- Blue teams receive user request and need to respond

Existing approaches

- GPT-style models integrated into support workflows

CYCON

Generative AI can generate reports

Problem

- Manual incident reporting is labor-intensive

Potential for generative AI

- LLMs can synthesize logs, timelines, and IoCs into structured, readable reports

Relevance in Locked Shields

- Blue teams are scored on the quality of various reports

Existing approaches

- GPT-based incident summarization (e.g., CYGENT)
- LLMs integrated with logging platforms

CYCON

Generative AI can efficiently integrate threat intelligence feeds and SIEM systems

Problem

- Manually correlating threat intelligence with logs is slow

Potential for generative AI

- LLMs can automate enrichment, prioritization, and action suggestions in SIEM workflows

Relevance in Locked Shields

- LS involves rapidly evolving threats and requires efficient information sharing

Existing approaches

- Integration with threat feeds, internal logs, and automated response mechanisms

CYCON

Generative AI offers many new opportunities

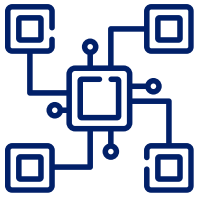
Automated patching

Ticket parsing

Network analysis

Report generation

Log analysis



Data integration

CYCON

Everything solved?

CYCON

Challenges

- Data availability
- Prompt engineering
- Integration complexity
- Computational resources
- Measuring effectiveness

CYCON

Challenges

- Data availability
- Prompt engineering
- Integration complexity
- Computational resources
- Measuring effectiveness

- High-quality, labeled datasets are scarce but critical for training and evaluating models
- LS datasets are a step forward, but real-world variation is still a gap

CYCON

Challenges

- Data availability
- Prompt engineering
- Integration complexity
- Computational resources
- Measuring effectiveness

- Effective LLM performance relies on carefully crafted prompts
- Automating prompt generation remains a challenge

CYCON

Challenges

- Data availability
- Prompt engineering
- Integration complexity
- Computational resources
- Measuring effectiveness

- LS and real-world environments consist of a large variety of different systems
- Seamless coordination of sensors, actuators, AI engines, and control logic is challenging

CYCON

Challenges

- Data availability
- Prompt engineering
- Integration complexity
- Computational resources
- Measuring effectiveness

- LLMs are resource-intensive
- Real-world use-cases might require running LLMs on premise

CYCON

Challenges

- Data availability
- Prompt engineering
- Integration complexity
- Computational resources
- Measuring effectiveness

- LLMs may fabricate plausible but incorrect outputs
- Standardized benchmarks and reproducible test environments are lacking
- LS provides a valuable setting, but frequency and scope are limited

CYCON

What should we do next?

CYCON

Next steps

- High-impact use-cases
- Reproducible test environments
- Data collection

CYCON

Next steps

- High-impact use-cases
- Reproducible test environments
- Data collection

- Automating support ticket processing
- Generating human-readable reports
- Detecting and fixing misconfigurations
- Combining data for actionable insights

CYCON

Next steps

- High-impact use-cases
- Reproducible test environments
- Data collection

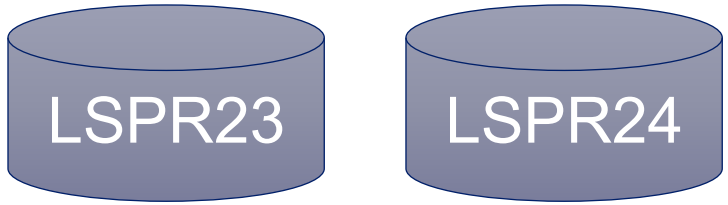
- Establish a testing environment that supports frequent, repeatable tests
- Integrate automated attacks as a counterpart to our defense techniques

CYCON

Next steps

- High-impact use-cases
- Reproducible test environments
- Data collection

- (Even) more data collection and labeling
- For example in future iterations of Locked Shields and in similar environments

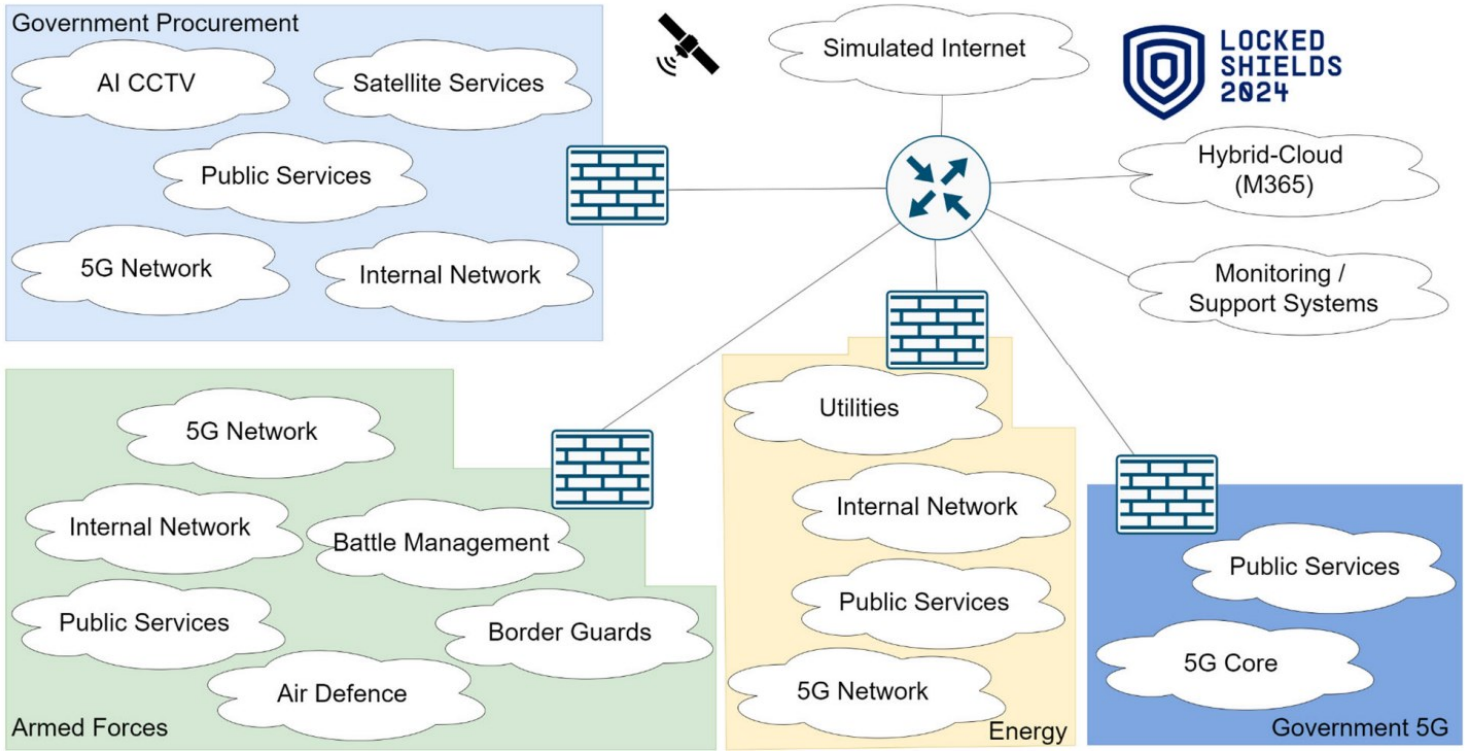


CYCON

We publish LSPR24, a dataset collected during the partners run of Locked Shields 2024

The dataset contains aggregated features and logs based on

- 400 GB network traffic
- 20 million flows
- 2 billion packets



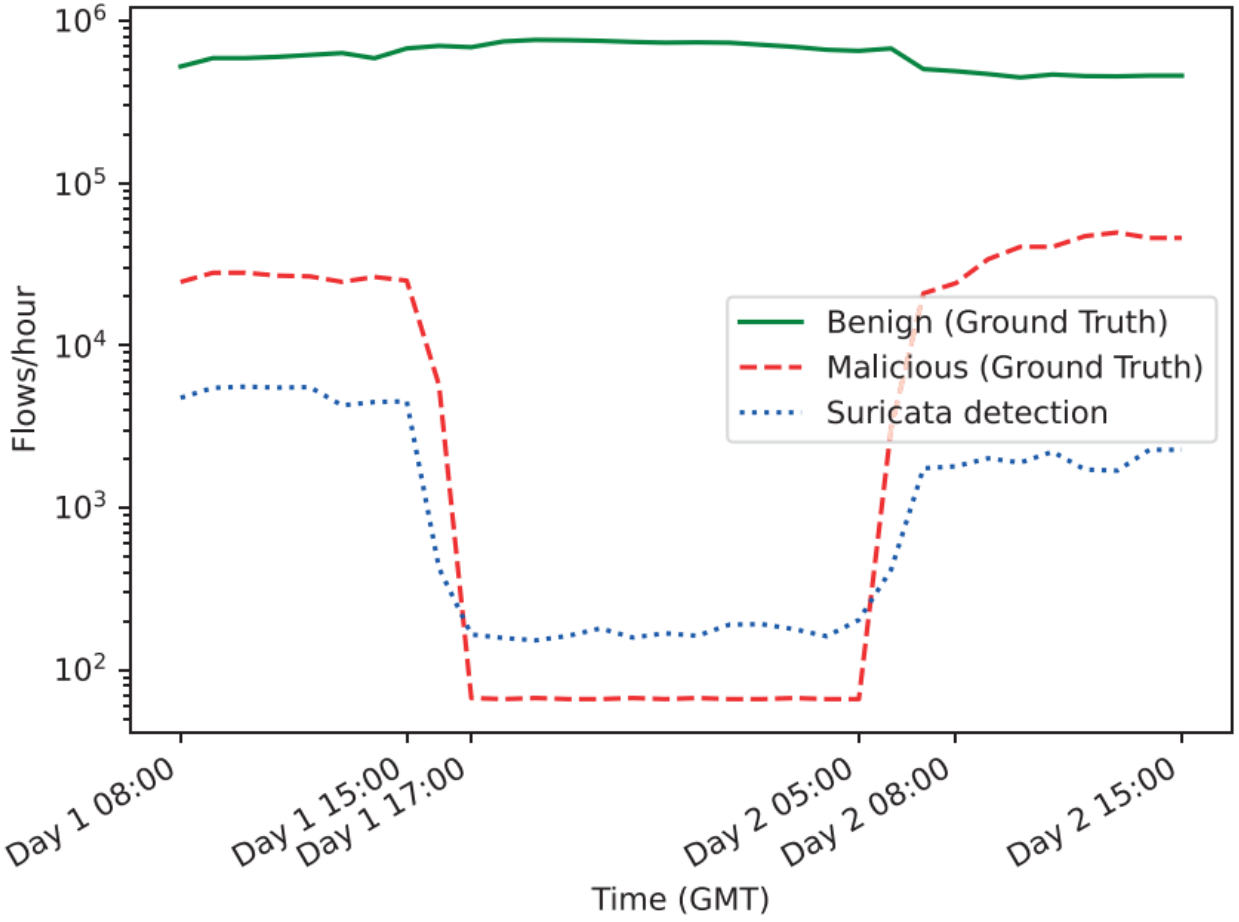
https://doi.org/10.5281/zenodo.14900873

CYCON

We publish LSPR24, a dataset collected during the partners run of Locked Shields 2024

The dataset contains aggregated features and logs based on

- 400 GB network traffic
- 20 million flows
- 2 billion packets



CYCON

Next Steps in Cyber Blue Team Automation — Leveraging the Power of LLMs

Generative AI offers many new possibilities to automate cyber defense

Remaining challenges include limited high-quality data, integration work and the need for benchmarking environments

To foster research, we publish the LSPR24 dataset

CYCON

Roland Meier
roland.meier@ar.admin.ch