Towards a fully automated Blue Team at Locked Shields

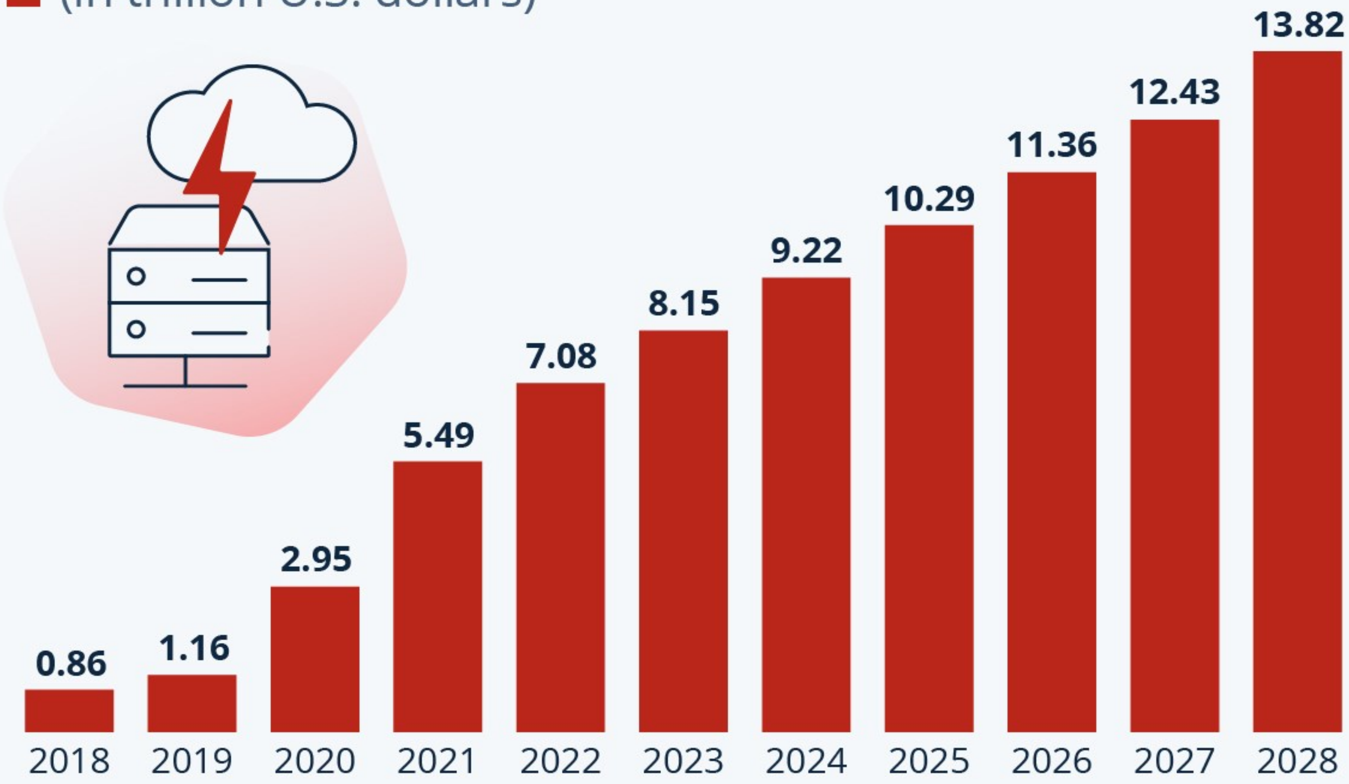
Dr. Roland Meier

armasuisse

Switzerland

Cybercrime Expected To Skyrocket

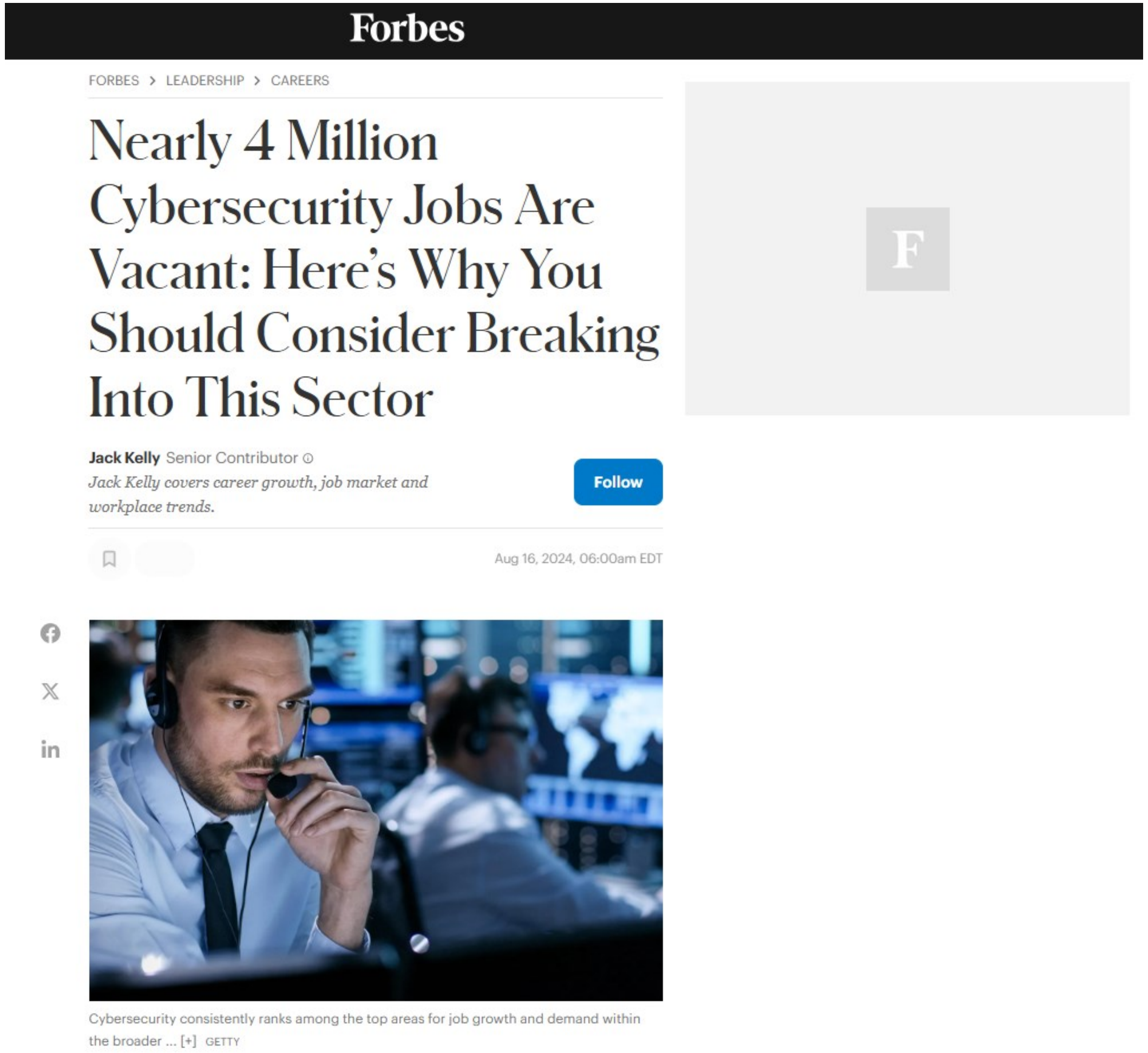
Estimated annual cost of cybercrime worldwide
(in trillion U.S. dollars)



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

Source: Statista Market Insights

There are not enough experts in this field



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.

Aug 16, 2024, 06:00am EDT

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

AI is everywhere…

THE REVOLUTIONARY GENIUS X WITH ARTIFICIAL INTELLIGENCE

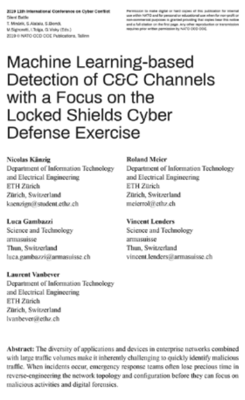
Recognizes your brushing style. Guides you to brush better every day.

... why not in cyber defense?

The HI behind the AI

- Allard Dijk
- Roland Meier
- Mauno Pihelgas CCDCOE
- Cosimo Melella CCDCOE
- Peter Hladký CCDCOE
- And many more

The history



2019 11th International Conference on Cyber Conflict

Machine Learning-based Detection of C&C Channels with a Focus on the Locked Shields Cyber Defense Exercise

Nicolas Känzig
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
kaenzign@student.ethz.ch

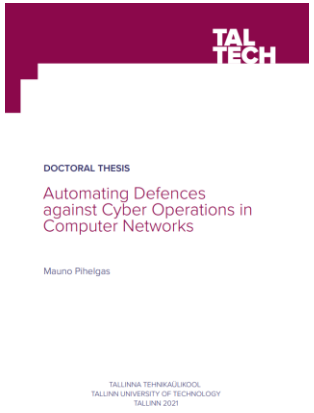
Roland Meier
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

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

Vincent Lenders
Science and Technology
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Laurent Vanbever
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
lvanbever@ethz.ch

Abstract: The diversity of applications and devices in enterprise networks combined with large traffic volumes make it inherently challenging to quickly identify malicious traffic. When incidents occur, emergency response teams often lose precious time in reverse-engineering the network topology and configuration before they can focus on malicious activities and digital forensics.



DOCTORAL THESIS

Automating Defences against Cyber Operations in Computer Networks

Mauno Pihelgas

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2021

Towards an Active, Autonomous and Intelligent Cyber Defense of Military Systems: the NATO AICA Reference Architecture

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

Arturs 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
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Abstract: Cyber attacks are becoming increasingly frequent, sophisticated, and stealthy. This makes it harder for cyber defence teams to keep up, forcing them to automate their defence capabilities in order to improve their reactivity and efficiency. Therefore, we propose a fully automated cyber defence framework that no longer needs support from humans to detect and mitigate attacks within a complex infrastructure. We design our framework based on a real-world case – Locked Shields – the world's largest cyber defence exercise. In this exercise, teams have to defend their networked infrastructure against attacks, while maintaining operational services for their users. Our framework architecture connects various cyber sensors with network, device, application, and user actuators through an artificial intelligence (AI)-powered automated team in order to dynamically secure the cyber environment. To the best of our knowledge, our framework is the first attempt towards a fully automated cyber defence team that aims at protecting complex environments from sophisticated attacks.

Keywords: *artificial intelligence, automation, Locked Shields, cyber defence, security*

2023 15th International Conference on Cyber Conflict

Towards Generalizing Machine Learning Models to Detect Command and Control Attack Traffic

Lina Gehri
ETH Zurich
Department of Electrical Engineering and Information Technology
Zurich, Switzerland
lina.gehri@gmail.com

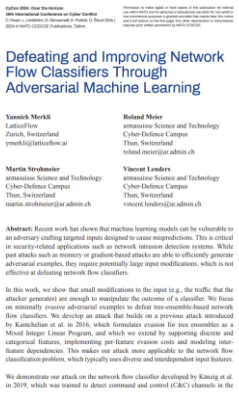
Roland Meier
Cyber-Defence Campus
armasuisse Science and Technology
Thun, Switzerland
roland.meier@ar.admin.ch

Daniel Hulliger
Cyber-Defence Campus
armasuisse Science and Technology
Thun, Switzerland
daniel.hulliger@ar.admin.ch

Vincent Lenders
Cyber-Defence Campus
armasuisse Science and Technology
Thun, Switzerland
vincent.lenders@ar.admin.ch

Abstract: Identifying compromised hosts from network traffic traces has become challenging because benign and malicious traffic is encrypted, and both use the same protocols and ports. Machine learning-based anomaly detection models have been proposed to address this challenge by classifying malicious traffic based on network flow features learned from historical patterns. Previous work has shown that such models successfully identify compromised hosts in the same network environment in which they were trained. However, cyber incidence response teams often have to look for intrusions in foreign networks, and we have found that learned models often fail to generalize to different network conditions. In this paper, we analyse the root cause of this problem using five network traces collected from different years and teams of Locked Shields, the world's largest live-fire cyber defence exercise. We then explore techniques to make machine learning models generalize better to unknown network environments and evaluate their accuracy.

Keywords: *machine learning, traffic classification, network security, command and control, Locked Shields*



CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict

Defeating and Improving Network Flow Classifiers Through Adversarial Machine Learning

Yannick Merkli
LatticeFlow
Zurich, Switzerland
ymerkli@latticeflow.ai

Roland Meier
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
roland.meier@ar.admin.ch

Martin Strohmeier
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
martin.strohmeier@ar.admin.ch

Vincent Lenders
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
vincent.lenders@ar.admin.ch

Abstract: Recent work has shown that machine learning models can be vulnerable to an adversary crafting targeted inputs designed to cause mispredictions. This is critical in security-related applications such as network intrusion detection systems. While past attacks such as mimicry or gradient-based attacks are able to efficiently generate adversarial examples, they require potentially large input modifications, which is not effective at defeating network flow classifiers.

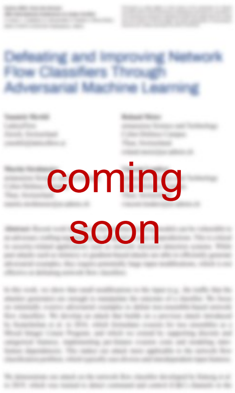
In this work, we show that small modifications to the input (e.g., the traffic that the attacker generates) are enough to manipulate the outcome of a classifier. We focus on minimally evasive adversarial examples to defeat tree-ensemble-based network flow classifiers. We develop an attack that builds on a previous attack introduced by Kantchelian et al. in 2016, which formulates evasion for tree ensembles as a Mixed Integer Linear Program, and which we extend by supporting discrete and categorical features, implementing per-feature evasion costs and modeling inter-feature dependencies. This makes our attack more applicable to the network flow classification problem, which typically uses diverse and interdependent input features.

We demonstrate our attack on the network flow classifier developed by Känzig et al. in 2019, which was trained to detect command and control (C&C) channels in the

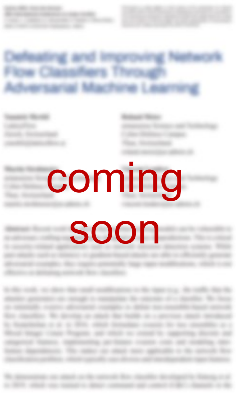
Journal of Information Security and Applications

LSPR23: A novel IDS dataset from the largest live-fire cybersecurity exercise

Allard Dijk, Emre Halisdemir, Cosimo Melella, Alari Schu, Mauno Pihelgas, Roland Meier



coming soon



coming soon

Research goals

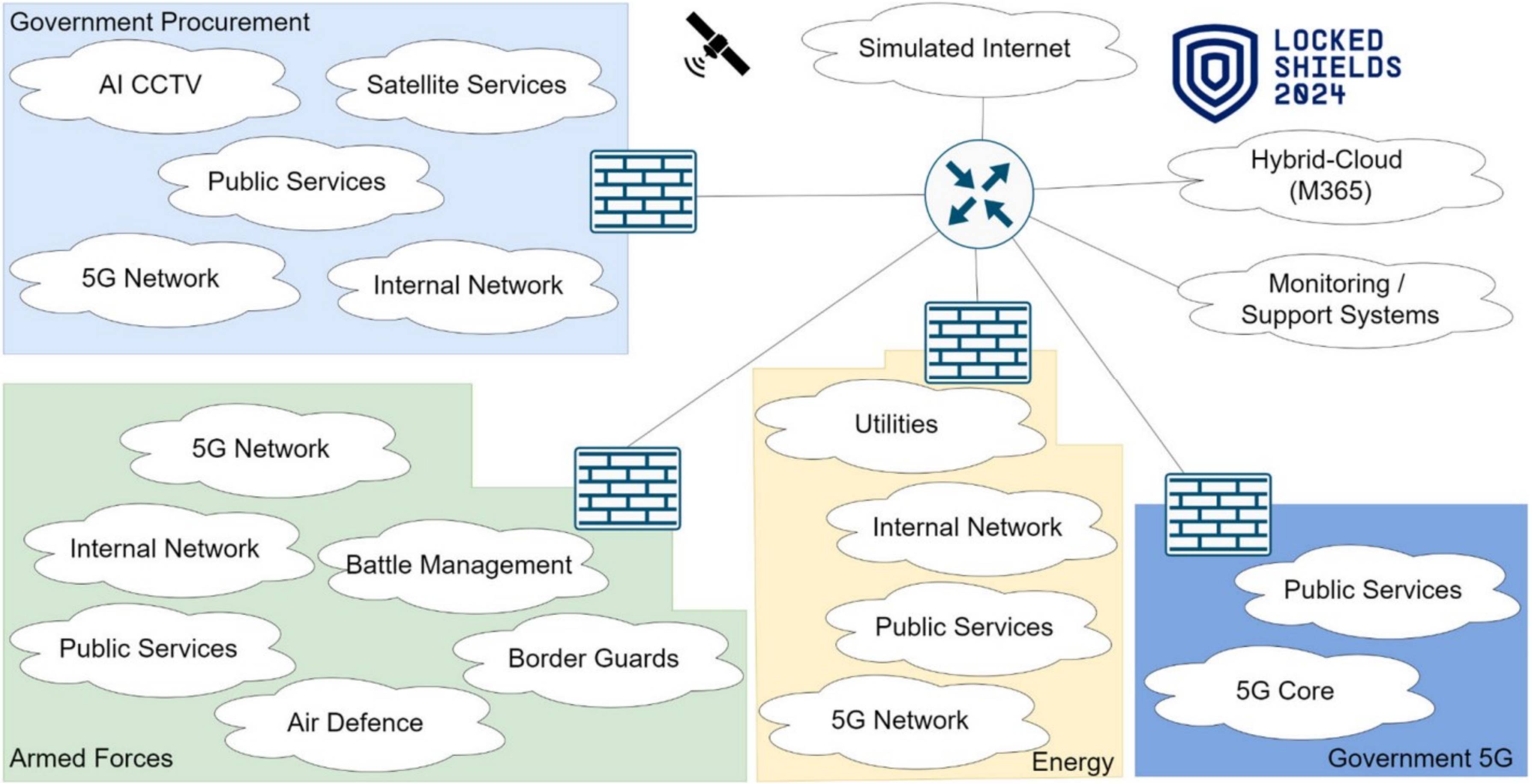
How can automation / AI help for cyber defense?

And eventually…

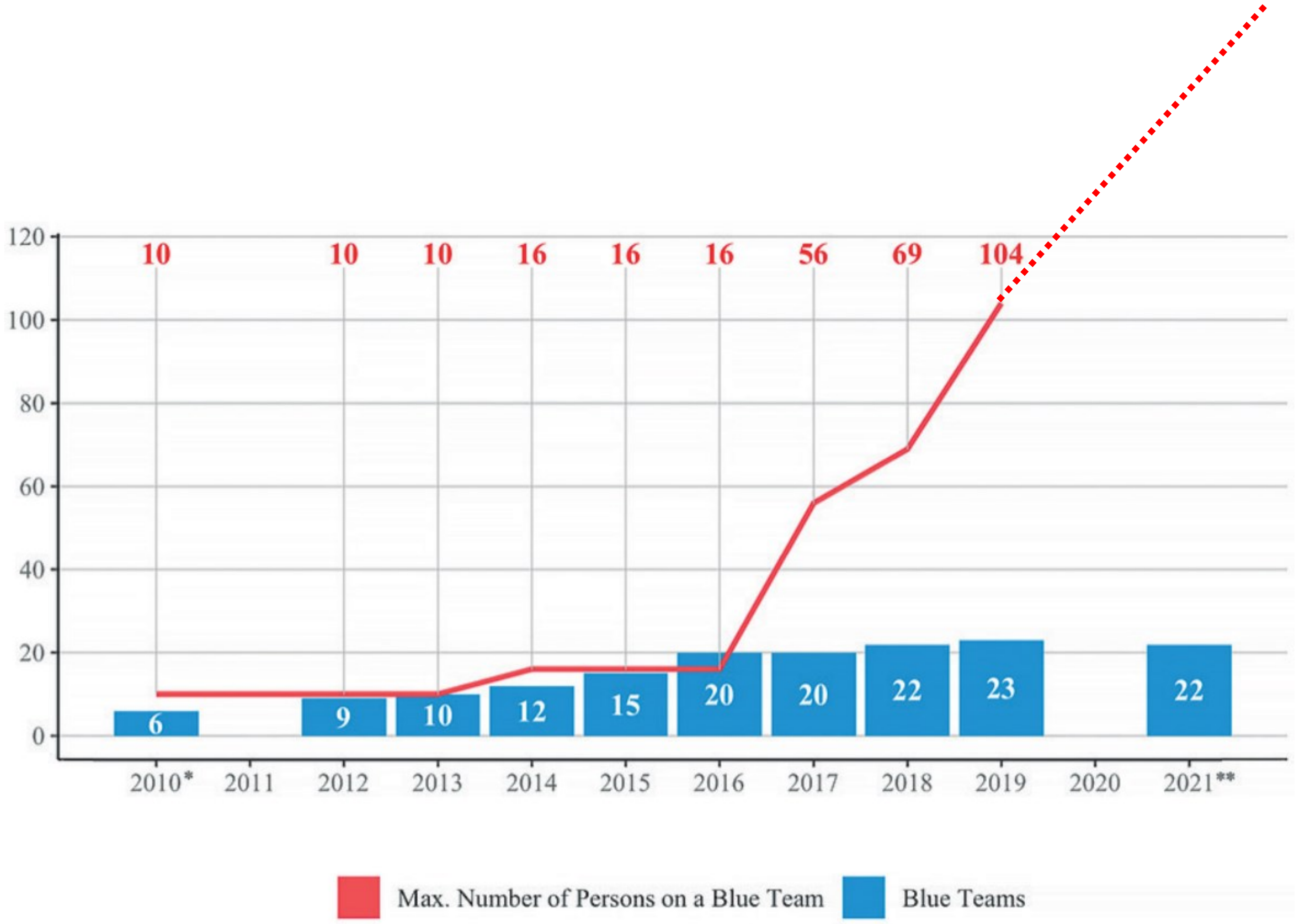
What would it take to have a fully automated Blue Team

in a future iteration of Locked Shields?

Locked Shields is the ideal testing ground for AI research



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*]

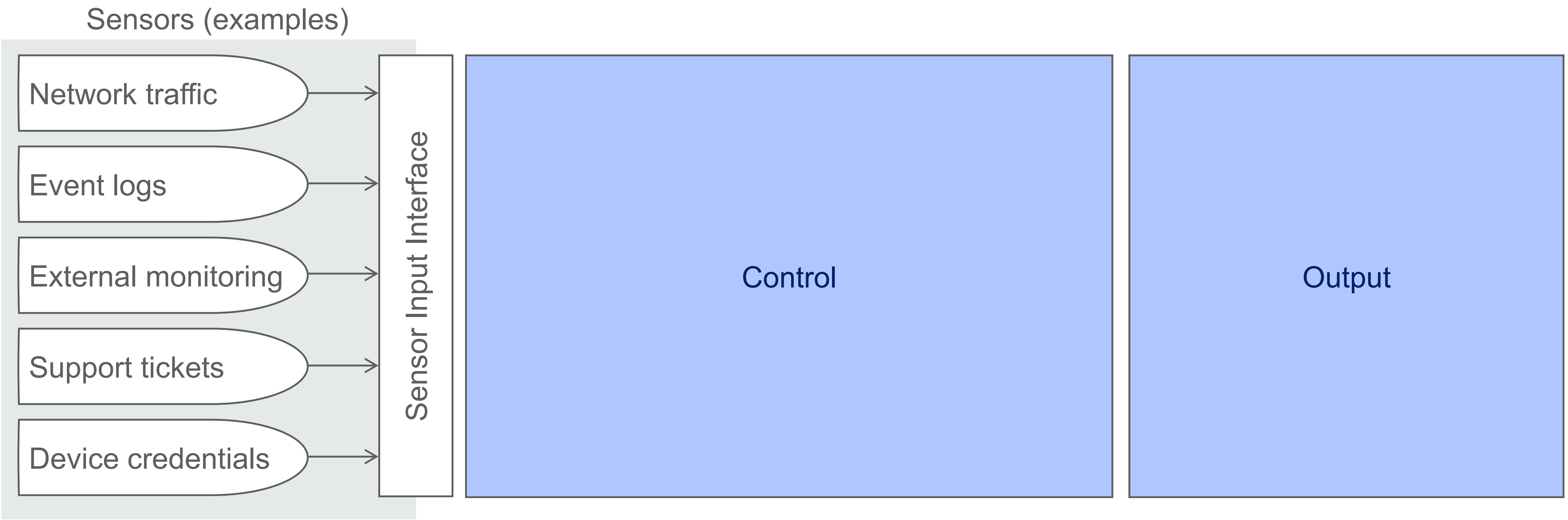
Automated Blue Team framework overview

Input

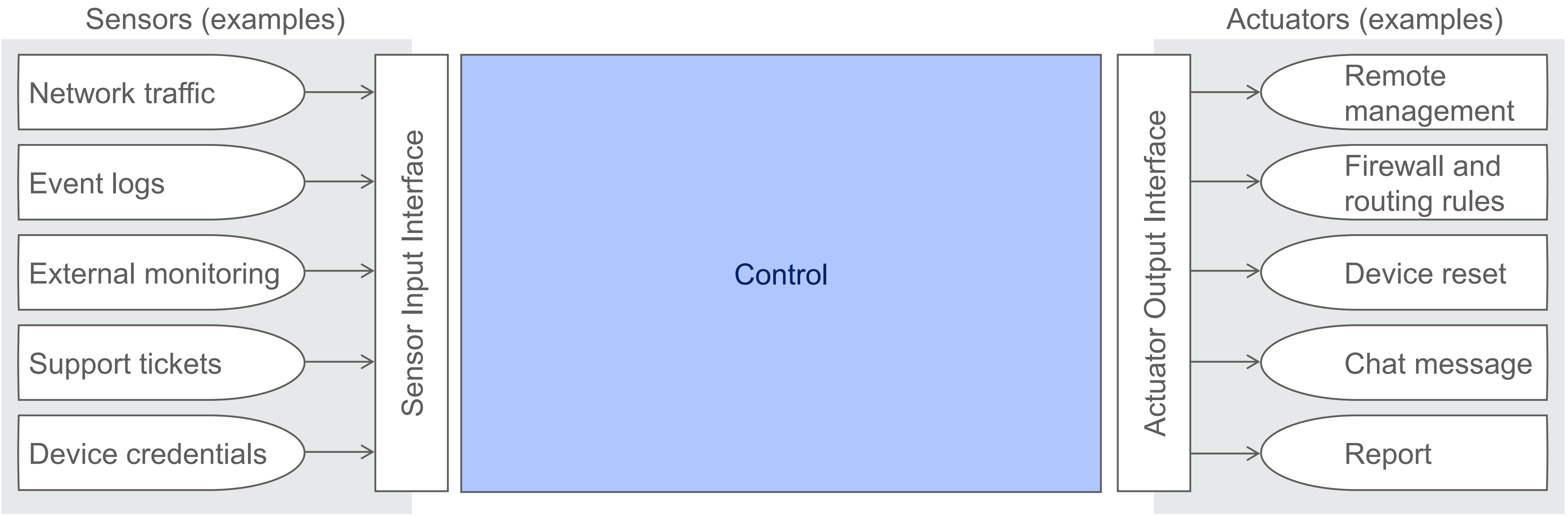
Control

Output

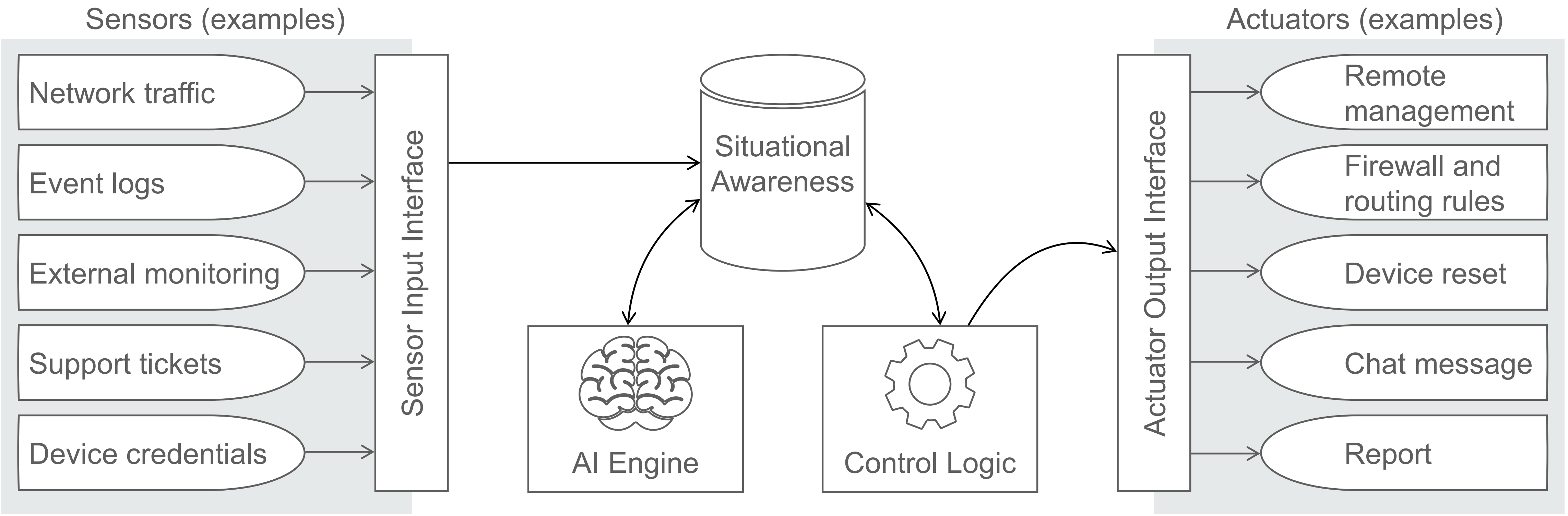
Automated Blue Team framework overview



Automated Blue Team framework overview



Automated Blue Team framework overview



Sensors (examples)

Network traffic

Event logs

External monitoring

Support tickets

Device credentials

Sensor Input Interface

Situational Awareness

AI Engine

Control Logic

Actuator Output Interface

Actuators (examples)

Remote management

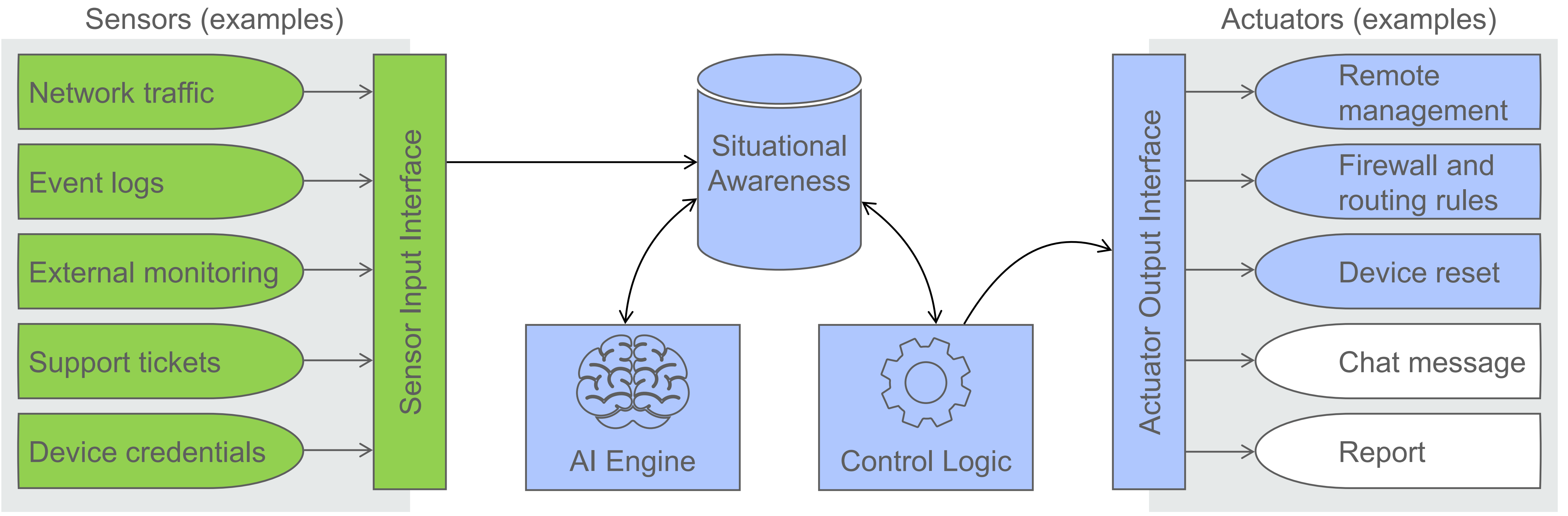
Firewall and routing rules

Device reset

Chat message

Report

First we focused on the data collection



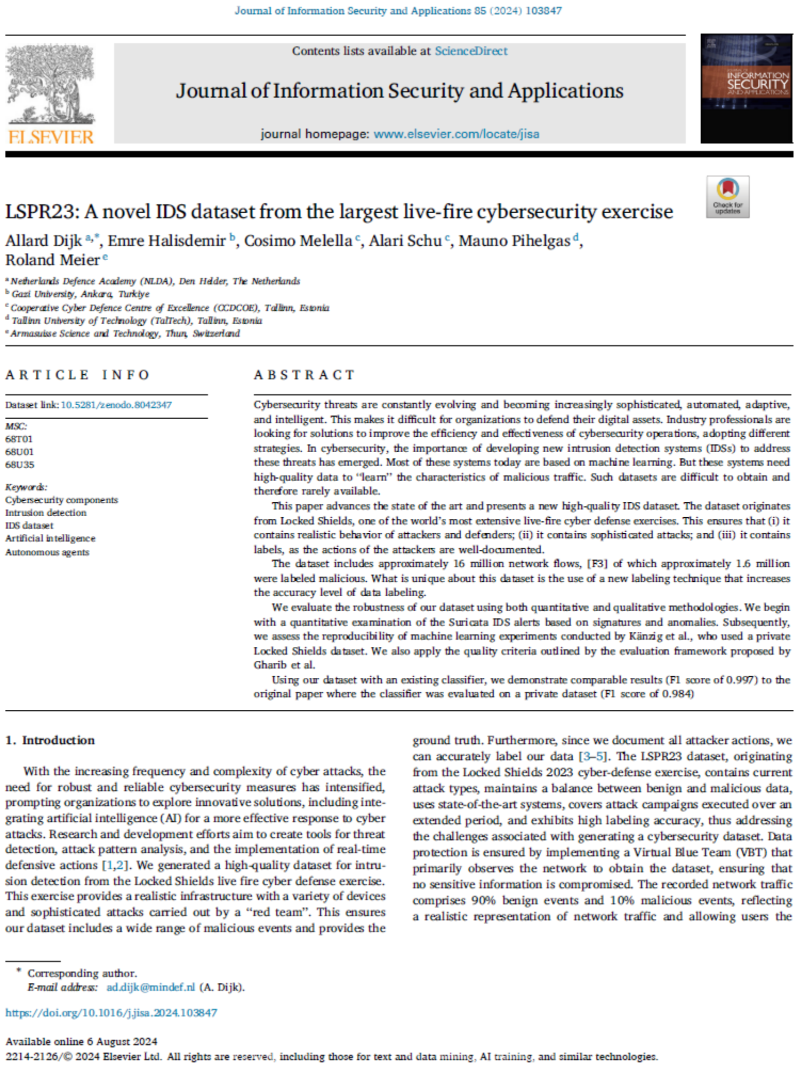
Since Locked Shields 2023 we use the partners run for data collection

- We participated as a blue team during the partners run
- Results include two public datasets

LSPR 23

LSPR 24

- Network traffic
- Labels
- Event logs



Journal of Information Security and Applications 85 (2024) 103847

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

LSPR23: A novel IDS dataset from the largest live-fire cybersecurity exercise

Allard Dijk [a,*], Emre Halisdemir [b], Cosimo Melella [c], Alari Schu [c], Mauno Pihelgas [d], Roland Meier [e]

[a] *Netherlands Defence Academy (NLDA), Den Helder, The Netherlands*
[b] *Gazi University, Ankara, Turkiye*
[c] *Cooperative Cyber Defence Centre of Excellence (CCDCOE), Tallinn, Estonia*
[d] *Tallinn University of Technology (TalTech), Tallinn, Estonia*
[e] *Armasuisse Science and Technology, Thun, Switzerland*

ARTICLE INFO

Dataset link: 10.5281/zenodo.8042347

MSC:
68T01
68U01
68U35

Keywords:
Cybersecurity components
Intrusion detection
IDS dataset
Artificial intelligence
Autonomous agents

ABSTRACT

Cybersecurity threats are constantly evolving and becoming increasingly sophisticated, automated, adaptive, and intelligent. This makes it difficult for organizations to defend their digital assets. Industry professionals are looking for solutions to improve the efficiency and effectiveness of cybersecurity operations, adopting different strategies. In cybersecurity, the importance of developing new intrusion detection systems (IDSs) to address these threats has emerged. Most of these systems today are based on machine learning. But these systems need high-quality data to "learn" the characteristics of malicious traffic. Such datasets are difficult to obtain and therefore rarely available.

This paper advances the state of the art and presents a new high-quality IDS dataset. The dataset originates from Locked Shields, one of the world's most extensive live-fire cyber defense exercises. This ensures that (i) it contains realistic behavior of attackers and defenders; (ii) it contains sophisticated attacks; and (iii) it contains labels, as the actions of the attackers are well-documented.

The dataset includes approximately 16 million network flows, [F3] of which approximately 1.6 million were labeled malicious. What is unique about this dataset is the use of a new labeling technique that increases the accuracy level of data labeling.

We evaluate the robustness of our dataset using both quantitative and qualitative methodologies. We begin with a quantitative examination of the Suricata IDS alerts based on signatures and anomalies. Subsequently, we assess the reproducibility of machine learning experiments conducted by Känzig et al., who used a private Locked Shields dataset. We also apply the quality criteria outlined by the evaluation framework proposed by Gharib et al.

Using our dataset with an existing classifier, we demonstrate comparable results (F1 score of 0.997) to the original paper where the classifier was evaluated on a private dataset (F1 score of 0.984)

1. Introduction

With the increasing frequency and complexity of cyber attacks, the need for robust and reliable cybersecurity measures has intensified, prompting organizations to explore innovative solutions, including integrating artificial intelligence (AI) for a more effective response to cyber attacks. Research and development efforts aim to create tools for threat detection, attack pattern analysis, and the implementation of real-time defensive actions [1,2]. We generated a high-quality dataset for intrusion detection from the Locked Shields live fire cyber defense exercise. This exercise provides a realistic infrastructure with a variety of devices and sophisticated attacks carried out by a "red team". This ensures our dataset includes a wide range of malicious events and provides the ground truth. Furthermore, since we document all attacker actions, we can accurately label our data [3–5]. The LSPR23 dataset, originating from the Locked Shields 2023 cyber-defense exercise, contains current attack types, maintains a balance between benign and malicious data, uses state-of-the-art systems, covers attack campaigns executed over an extended period, and exhibits high labeling accuracy, thus addressing the challenges associated with generating a cybersecurity dataset. Data protection is ensured by implementing a Virtual Blue Team (VBT) that primarily observes the network to obtain the dataset, ensuring that no sensitive information is compromised. The recorded network traffic comprises 90% benign events and 10% malicious events, reflecting a realistic representation of network traffic and allowing users the

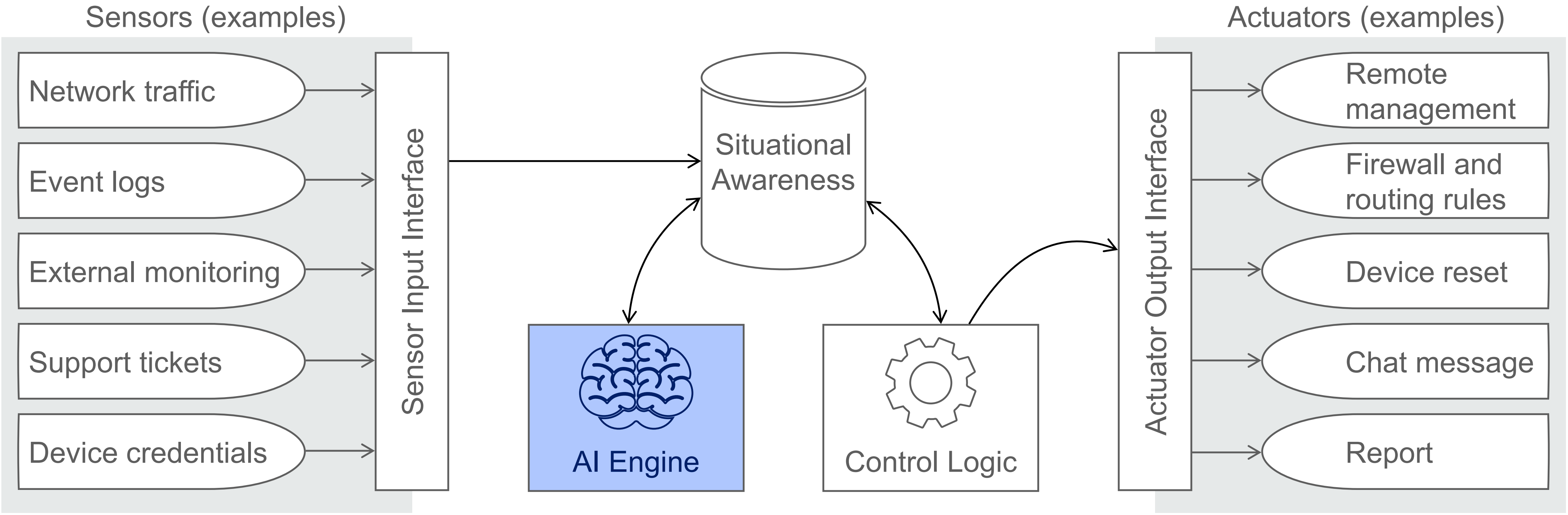
---

[*] Corresponding author.
E-mail address: ad.dijk@mindef.nl (A. Dijk).

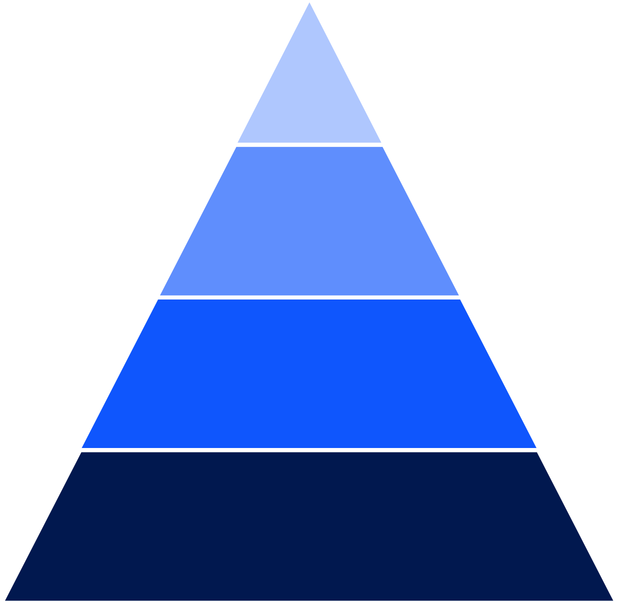
https://doi.org/10.1016/j.jisa.2024.103847

Available online 6 August 2024
2214-2126/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Automated Blue Team framework overview



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” today
Level 0: Reactive narrow AI	Signature-based decisions

We are also looking at Generative AI

Automated patching

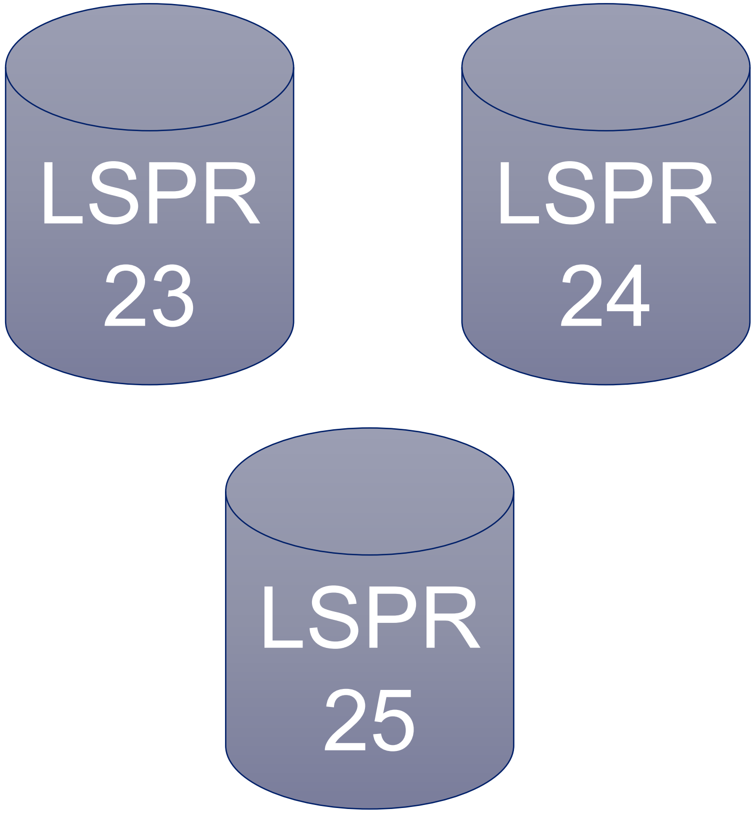
Network analysis

Log analysis

Ticket parsing

Report generation

And after the Locked Shields 2025?

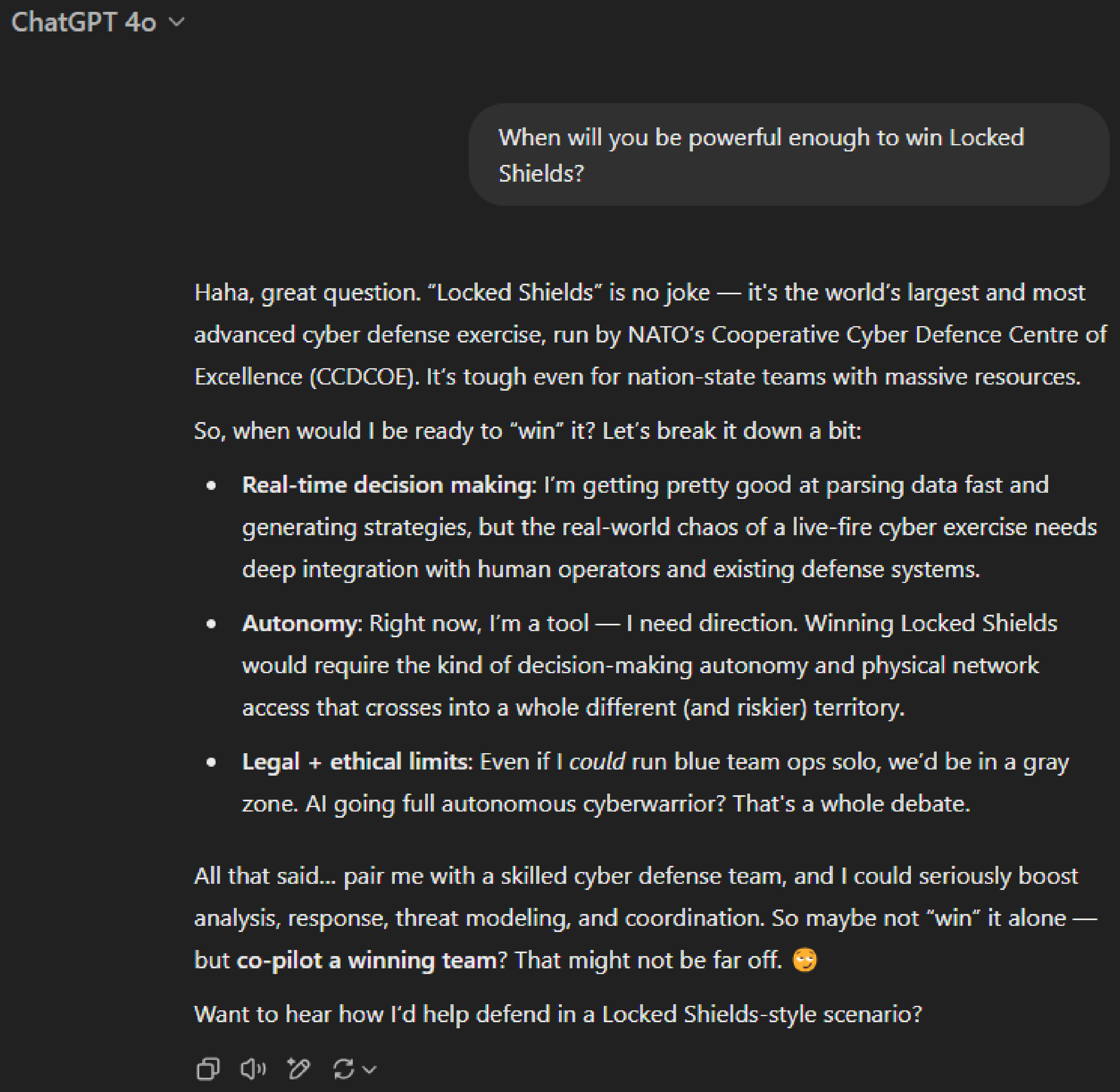


LSPR 23

LSPR 24

LSPR 25

- Automated detection and response to some types of attacks
- Applying LLMs for detecting and responding to attacks
- Evaluation in LS 2026?



When will you be powerful enough to win Locked Shields?

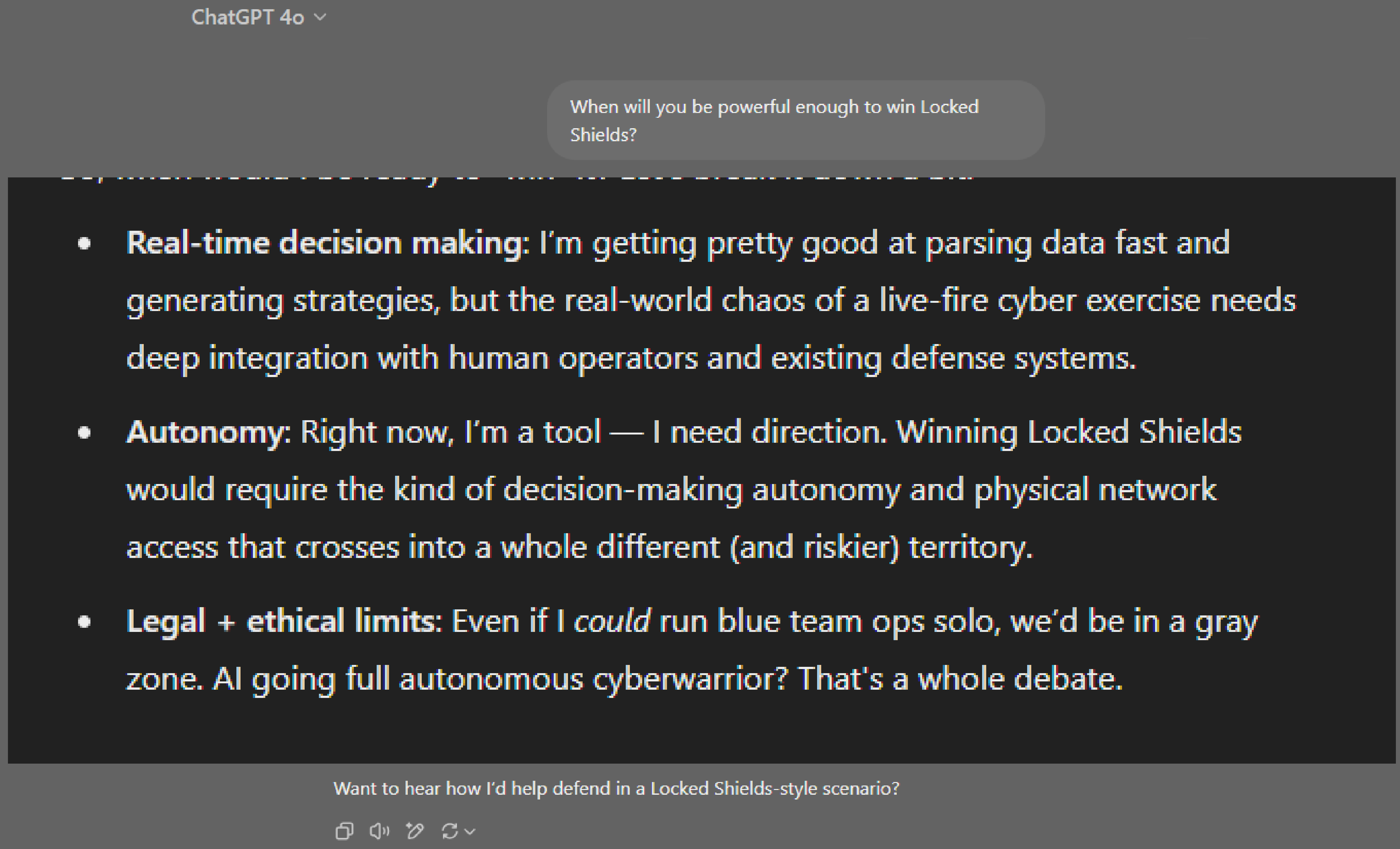
Haha, great question. “Locked Shields” is no joke — it's the world’s largest and most advanced cyber defense exercise, run by NATO’s Cooperative Cyber Defence Centre of Excellence (CCDCOE). It’s tough even for nation-state teams with massive resources.

So, when would I be ready to “win” it? Let’s break it down a bit:

- **Real-time decision making**: I’m getting pretty good at parsing data fast and generating strategies, but the real-world chaos of a live-fire cyber exercise needs deep integration with human operators and existing defense systems.
- **Autonomy**: Right now, I’m a tool — I need direction. Winning Locked Shields would require the kind of decision-making autonomy and physical network access that crosses into a whole different (and riskier) territory.
- **Legal + ethical limits**: Even if I *could* run blue team ops solo, we’d be in a gray zone. AI going full autonomous cyberwarrior? That's a whole debate.

All that said… pair me with a skilled cyber defense team, and I could seriously boost analysis, response, threat modeling, and coordination. So maybe not “win” it alone — but **co-pilot a winning team**? That might not be far off. 😌

Want to hear how I’d help defend in a Locked Shields-style scenario?



When will you be powerful enough to win Locked Shields?

- **Real-time decision making:** I'm getting pretty good at parsing data fast and generating strategies, but the real-world chaos of a live-fire cyber exercise needs deep integration with human operators and existing defense systems.
- **Autonomy:** Right now, I'm a tool — I need direction. Winning Locked Shields would require the kind of decision-making autonomy and physical network access that crosses into a whole different (and riskier) territory.
- **Legal + ethical limits:** Even if I *could* run blue team ops solo, we'd be in a gray zone. AI going full autonomous cyberwarrior? That's a whole debate.

Want to hear how I'd help defend in a Locked Shields-style scenario?

Towards a fully automated Blue Team at Locked Shields

Thanks for your attention!

Dr. Roland Meier

armasuisse