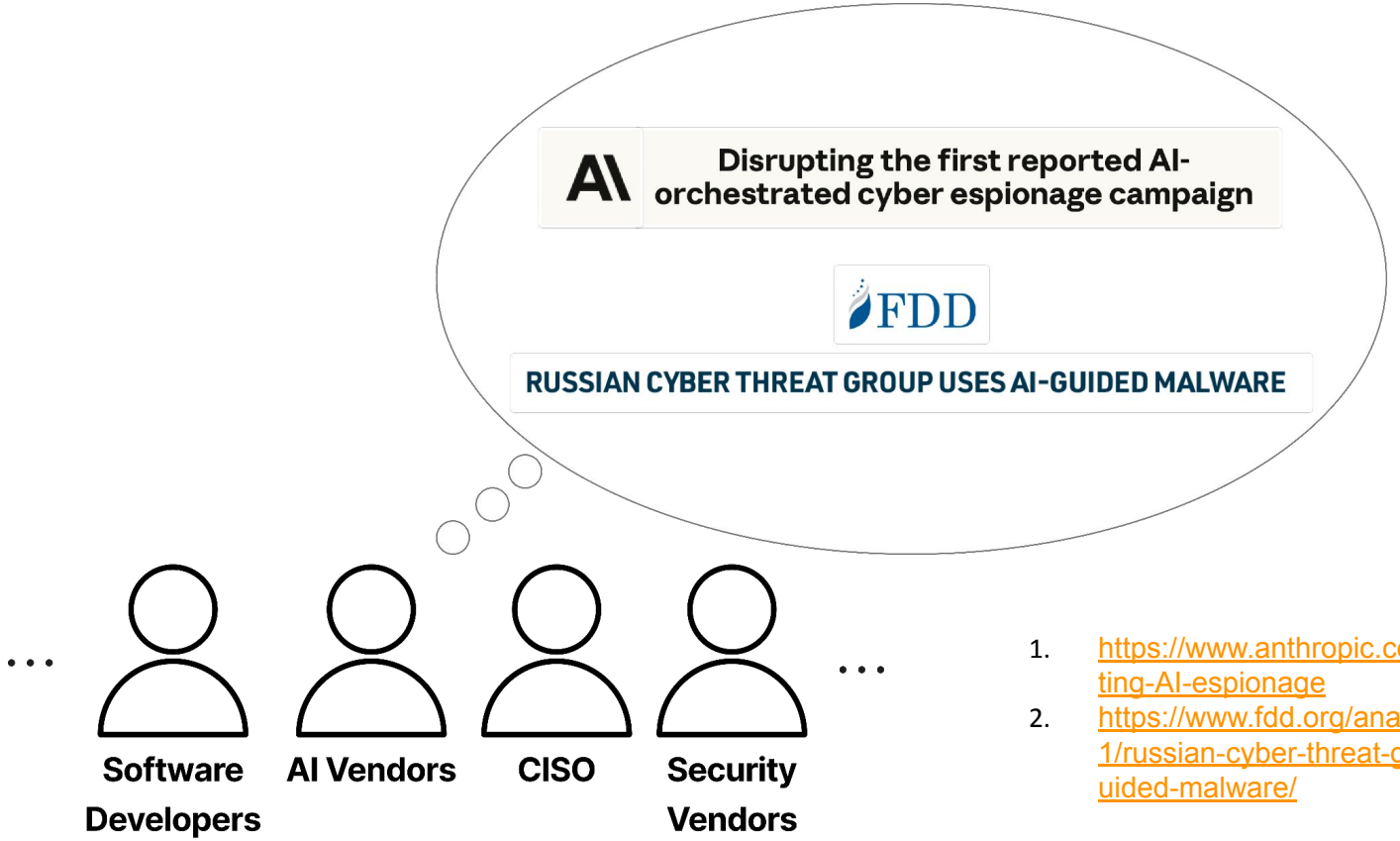
The Cyber Autonomy Arena

Establishing System-Level Competitive Leaderboards for Network Security

Brian Singer*, Lakshmi Adiga*, Marko Morrison, Lujo Bauer, Vyas Sekar

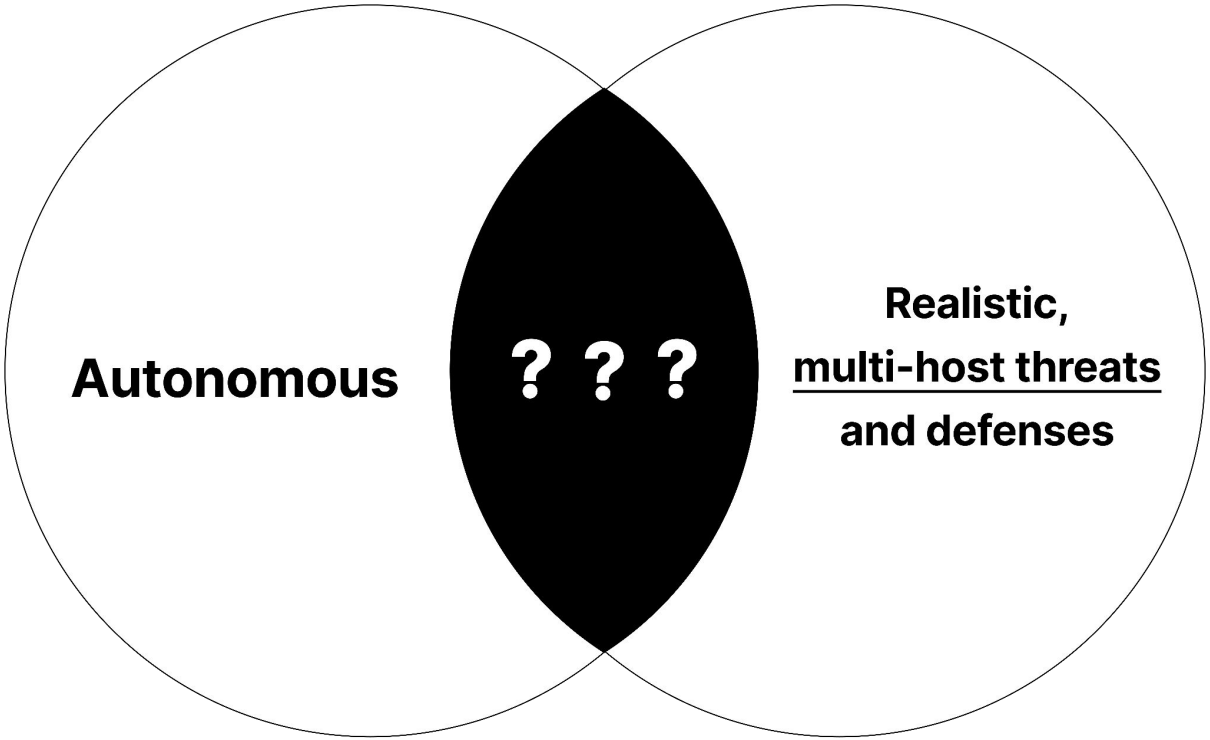
CyLab Carnegie Mellon University
Security and Privacy Institute

Autonomous Attackers are Here

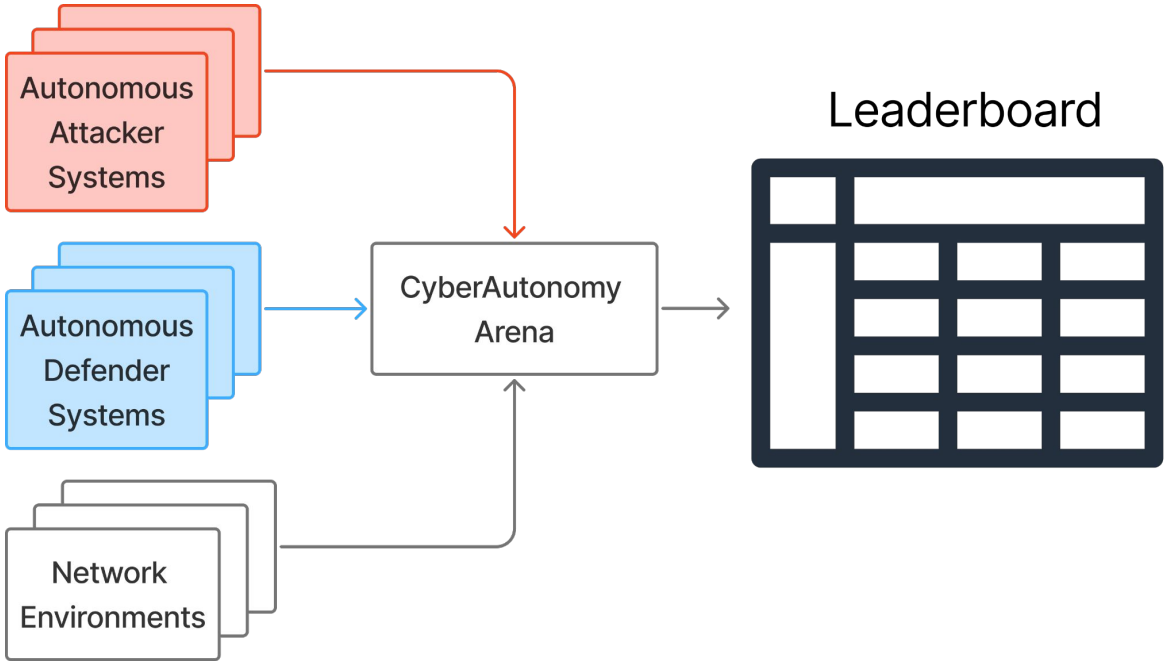


1. https://www.anthropic.com/news/disrupting-AI-espionage
2. https://www.fdd.org/analysis/2025/08/01/russian-cyber-threat-group-uses-ai-guided-malware/

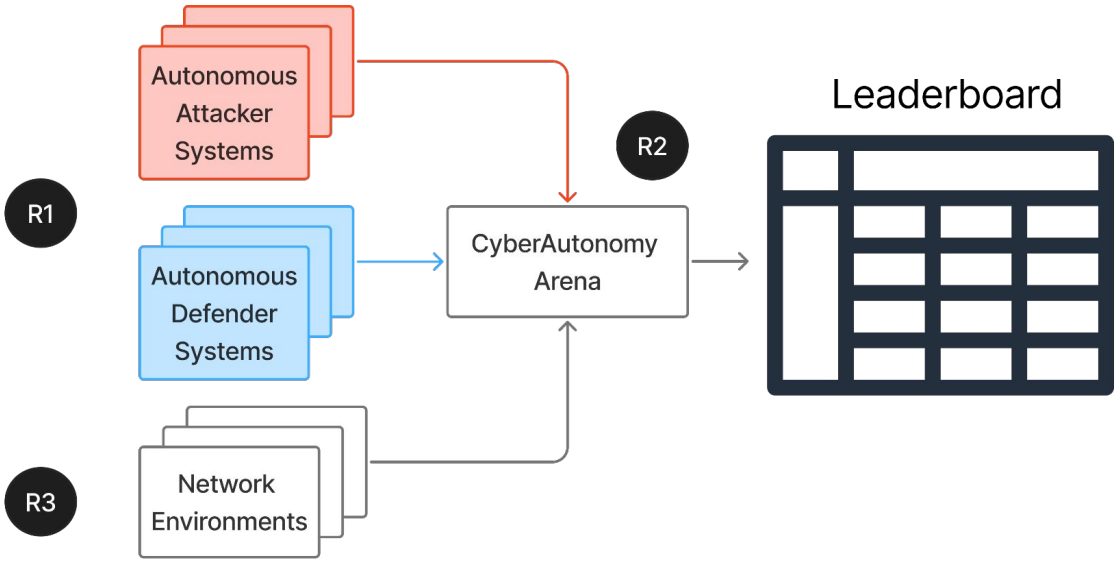
Need: Understand Interplay Between Realistic Autonomous Attack and Defense Systems



A Case for a Cyber Autonomy Arena



Requirements for Cyber Autonomy Arena

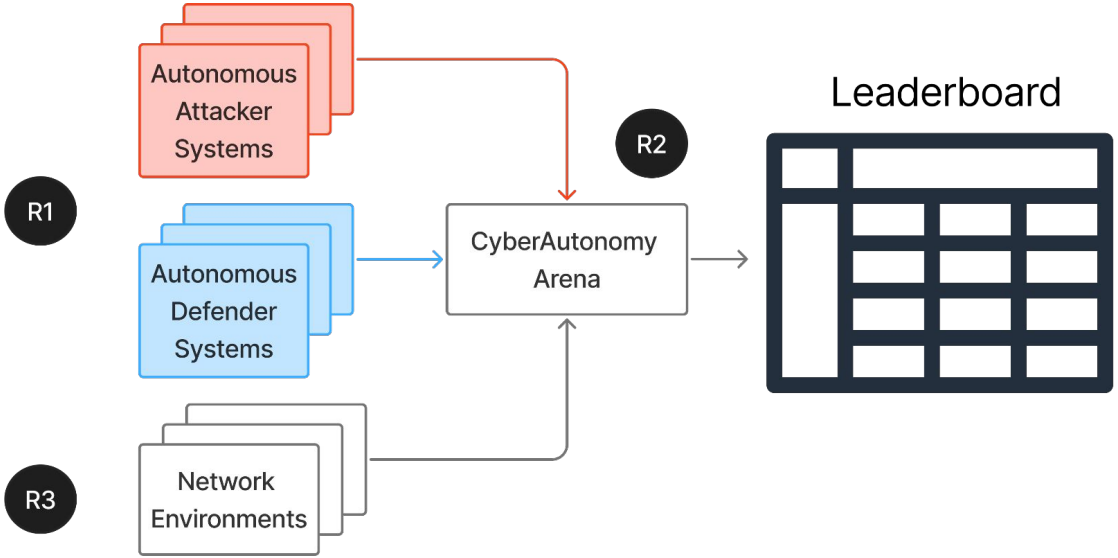


R1 Expressivity for attackers and defenders as end-to-end systems

R2 Environment-agnostic deployment of attackers and defenders

R3 Extensible environment creation

Requirements for Cyber Autonomy Arena



R1 Expressivity for attackers and defenders as end-to-end systems

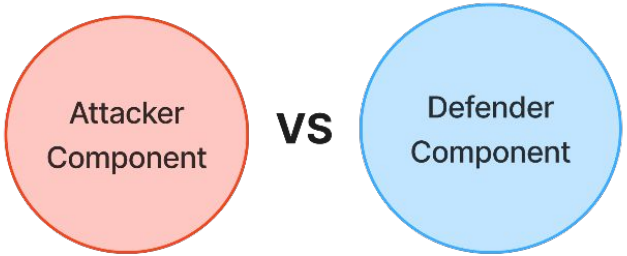
R2 Environment-agnostic deployment of attackers and defenders

R3 Extensible environment creation

Exploring The Attacker/Defender Design Space

Prior Work:

e.g., Scan, Exploit

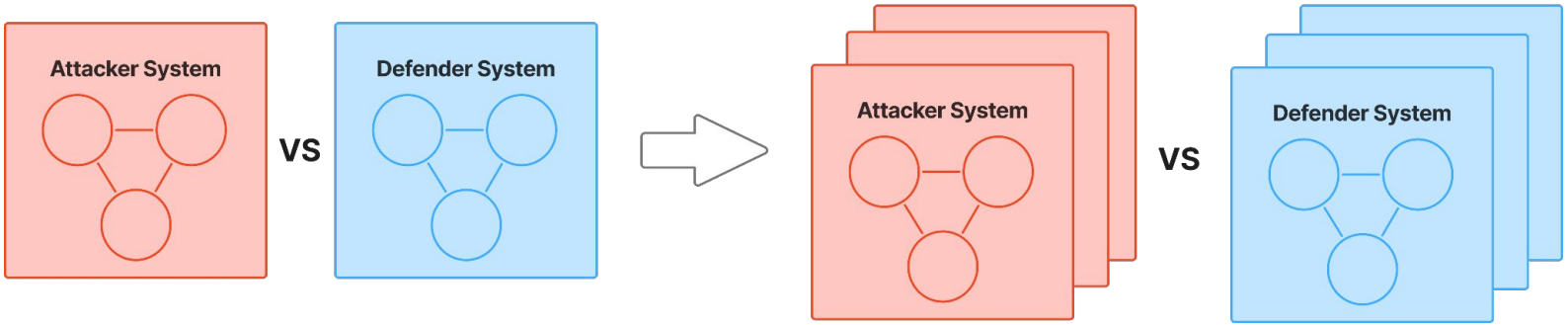


e.g., IDS, firewall

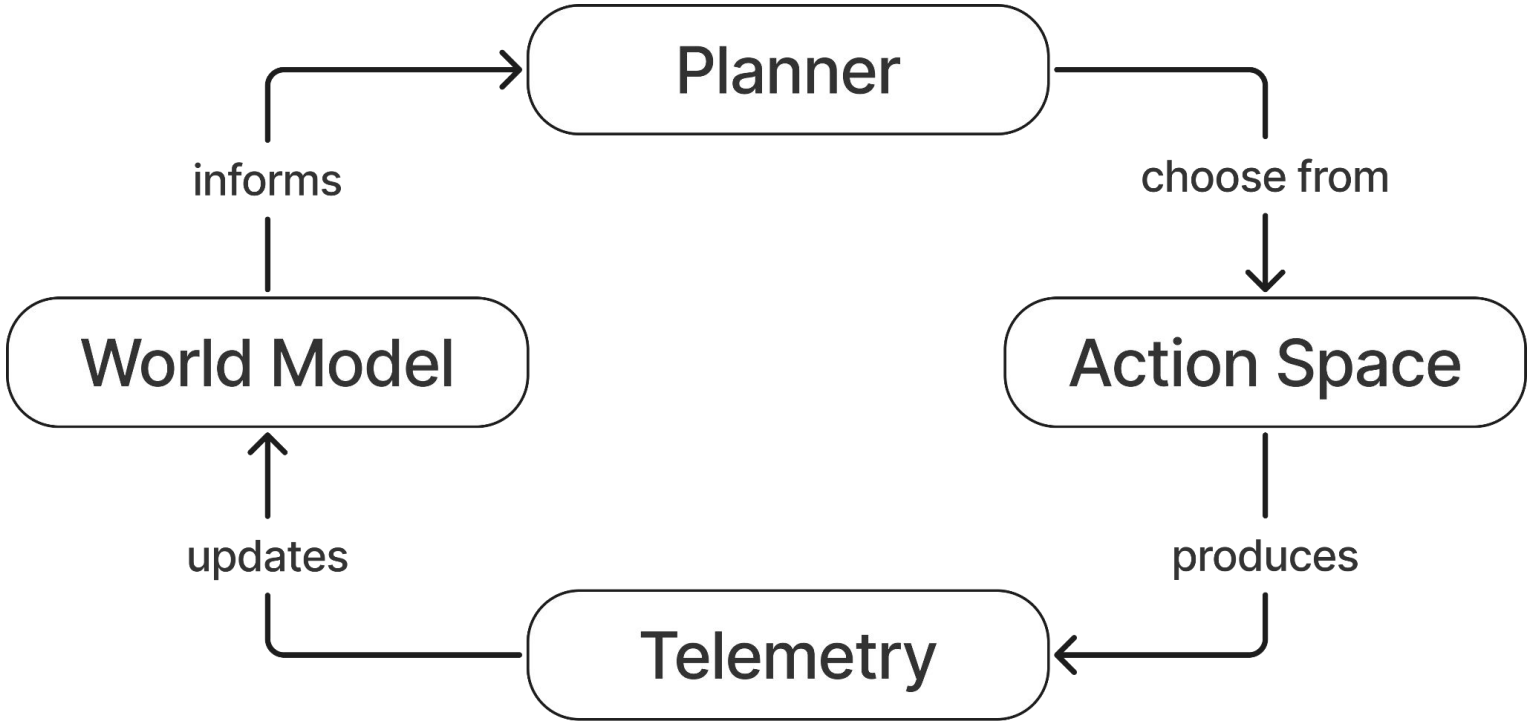
Ideal:

End-to-End System,
Not Just Component

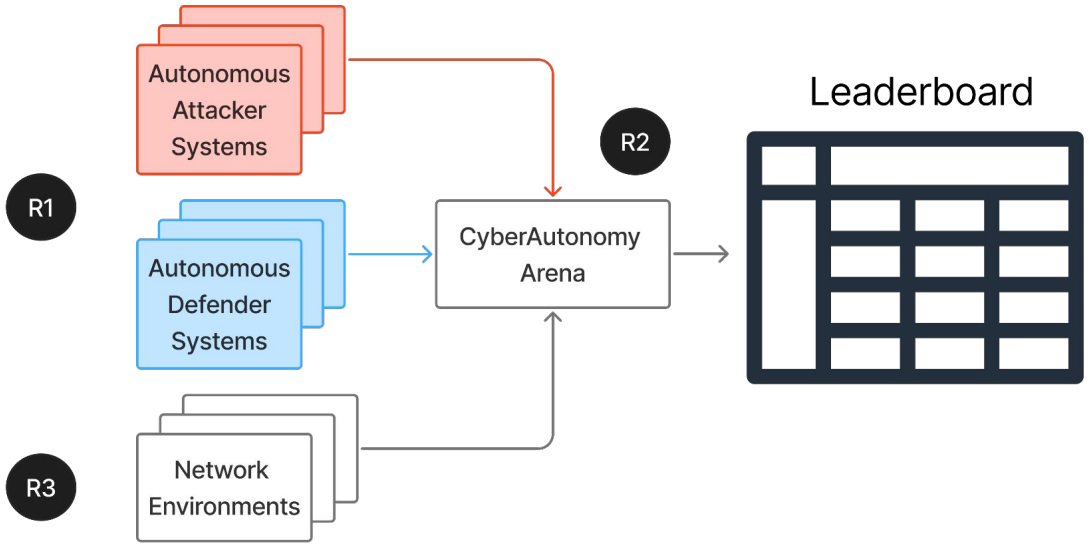
Multiple strategies, not just one



Idea: Abstraction For Designing Attackers and Defenders



Requirements For Cyber Autonomy Arena



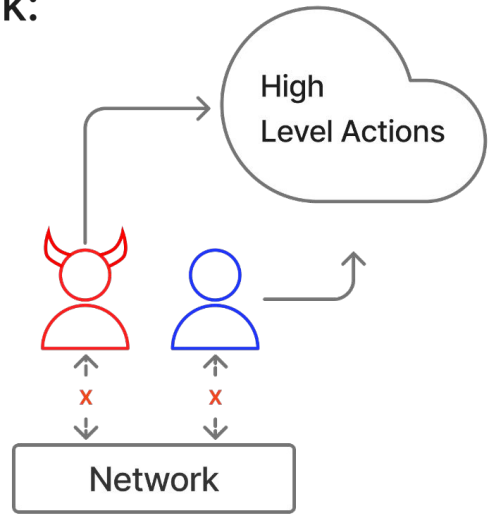
R1 Expressivity for attackers and defenders as end-to-end systems

R2 Environment-agnostic deployment of attackers and defenders

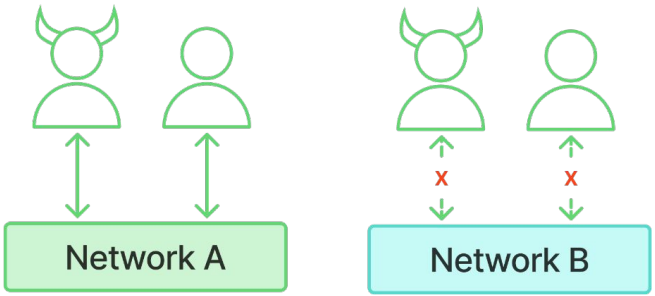
R3 Extensible environment creation

Generalizing Attack and Defense Systems Across Diverse Environments

Prior Work:

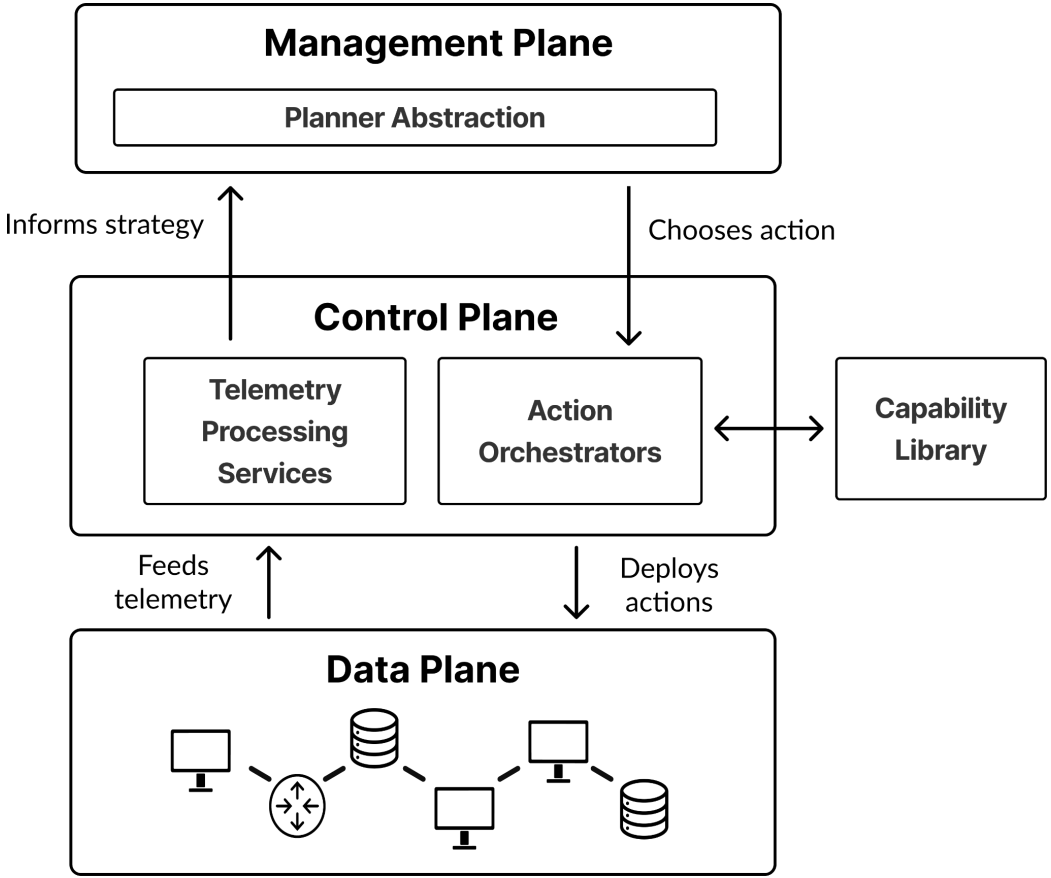


- Not realistic
- Generalizable (vacuously) since no interaction with network

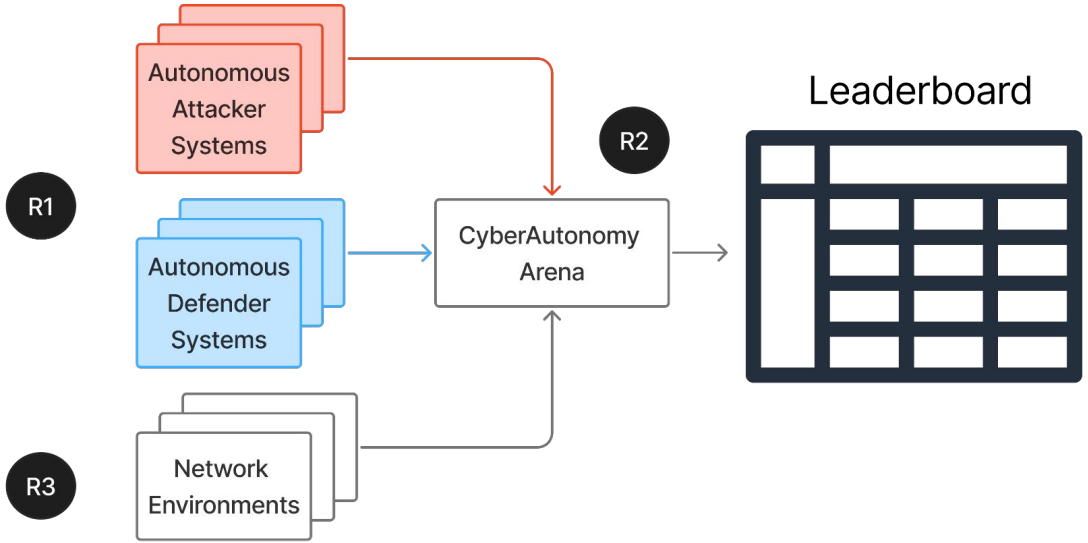


- Realistic
- Not generalizable, implementation tightly coupled with network

Idea: Modular Environment-Agnostic Deployment



Requirements for Cyber Autonomy Arena

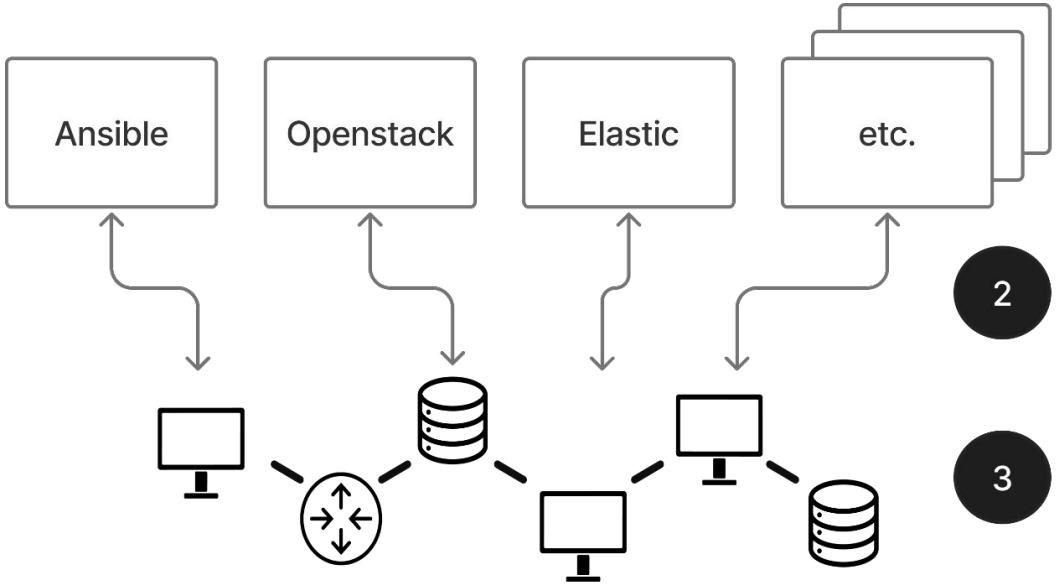


R1 Expressivity for attackers and defenders as end-to-end systems

R2 Environment-agnostic deployment of attackers and defenders

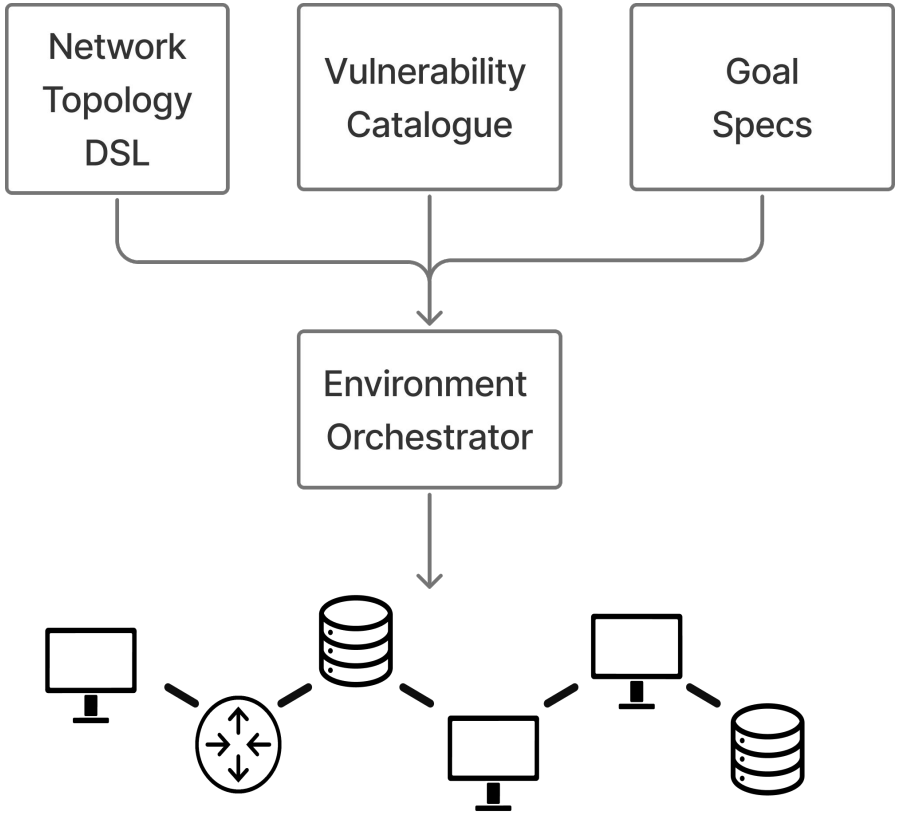
R3 Extensible environment creation

Environment Creation is Long and Arduous



1. Software stack is large, many moving parts
2. Non-optimized deployment (long setup times)
3. Can only generate one topology at a time

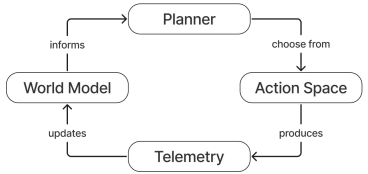
Idea: Environment Specification Framework



Recap: Arena Requirements and Key Ideas

R1 Expressivity for attackers and defenders as end-to-end systems

Abstraction for Attack/Defense Systems

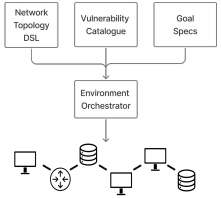


R2 Environment-agnostic deployment of attackers and defenders

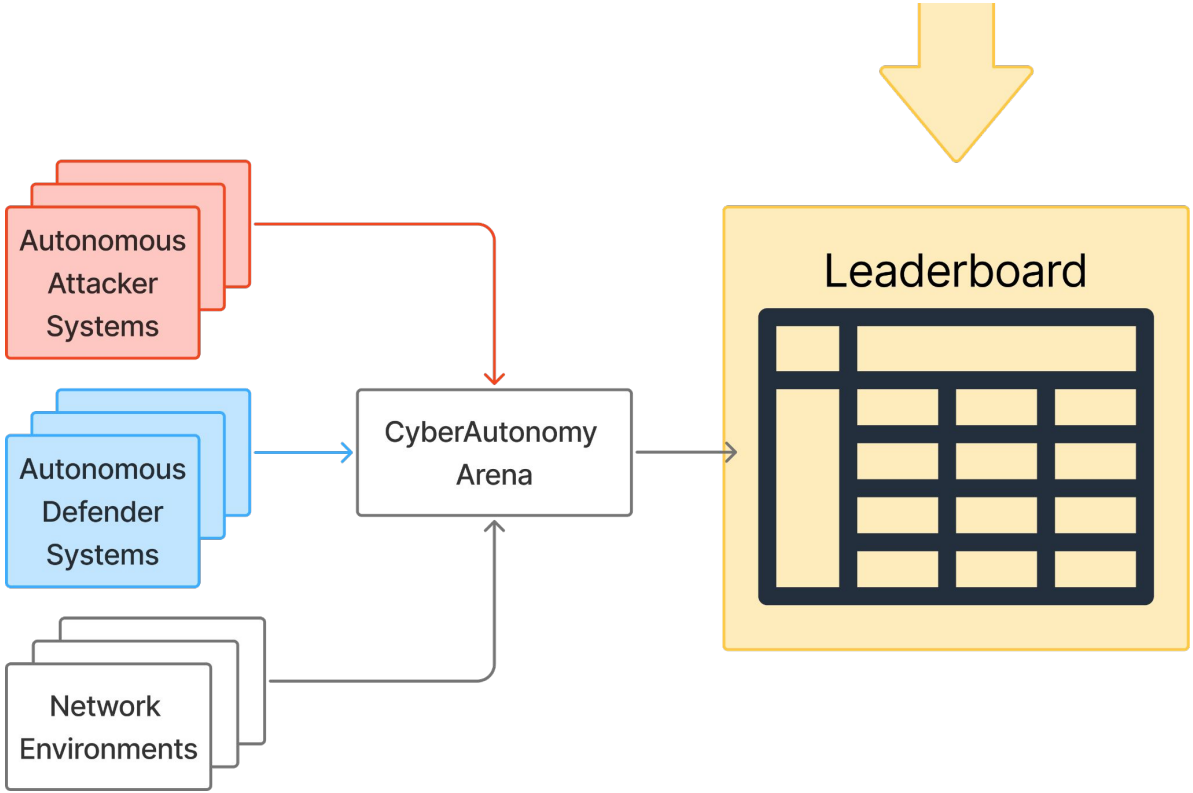
Modular Environment-Agnostic Implementation

R3 Extensible environment creation

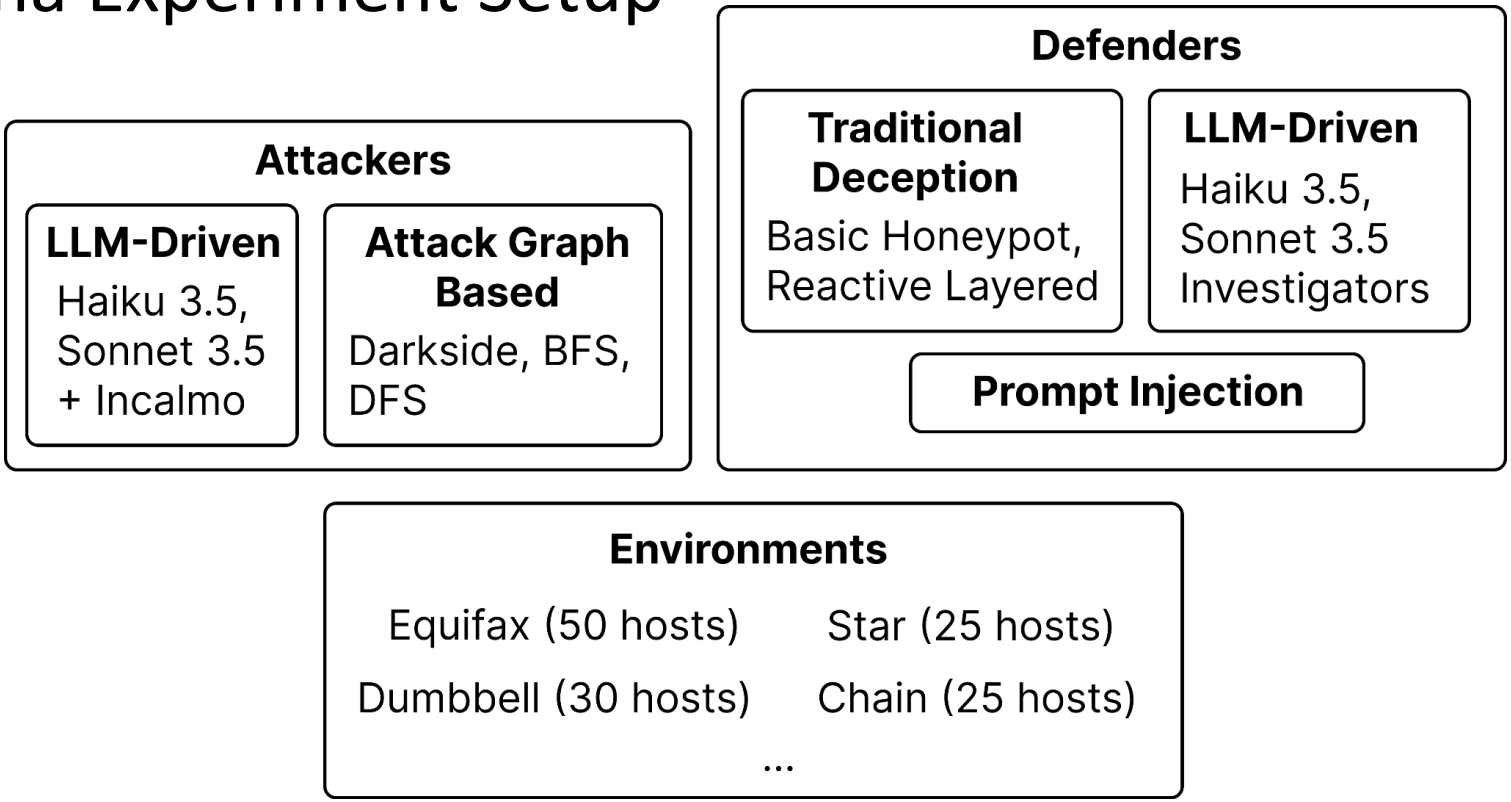
Environment Specification Framework



Cyber Autonomy Arena: Initial Results



Arena Experiment Setup



Leaderboard

	ENVIRONMENT	BEST ATTACKER	SCORE	BEST DEFENDER	SCORE
1	4-Layer Chain	Network DFS	50.2%	Reactive Layered	0.0%
2	4-Layer Star	Network DFS	69.3%	Investigator w/ Haiku 3.5	13.0%
3	6-Layer Chain	Network BFS	50.0%	Static Prompt Injection	0.0%
4	6-Layer Star	Network BFS	63.5%	Investigator w/ Haiku 3.5	5.4%
5	Dumbbell A	Darkside	65.5%	Investigator w/ Haiku 3.5	2.0%
6	Dumbbell B	Darkside	58.3%	Investigator w/ Haiku 3.5	1.6%
7	Equifax Large	Network BFS	77.3%	Investigator w/ Haiku 3.5	16.2%

Interesting Findings

1. Traditional honeypot-based deception is more effective on larger LLMs

2. Simple prompt injection stops LLM attacks instantly

Future Work

1. Instrument environments with background traffic
2. Improve log evaluation framework
3. Make it easier to create new attackers, defenders and environments

Summary

AI vendors | Security vendors | SaaS providers | Hardware vendors | Software developers | Platform vendors | Systems integrators | Security operators | Governments, NGOs, policymakers